TAMPERE UNIVERSITY OF TECHNOLOGY

VILLE KYTÖLÄ

COMPUTATIONAL PREDICTION OF PROSTATE CANCER SPECIFIC GENE REGULATORY NETWORK

Masters Thesis

Examiner: Professor Olli Yli-Harja
Examiner and topic approved in
the Faculty of Natural Sciences
Council meeting on 07.05.2014

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Biotekniikan koulutusohjelma
VILLE KYTÖLÄ: Eturauhassyöpäspesifisen geenisäätelyverkon laskennallinen ennustaminen
Diplomityö, 70 sivua, 0 liitesivua
Kesäkuu 2014
Pääaine: Laskennallinen systeemibiologia
Tarkastajat: Olli Yli-Harja
Avainsanat: Geenisäätelyverkko, Laskennallinen ennustaminen, Eturauhassyöpä

Eturauahssyöpä on länsimaissa hyvin yleinen sairaus, jonka esiintymistiheys kasvaa vuosi vuodelta elintason ja elinajanodotteen kanssa. Laajoista tutkimuksista huolimatta eturauhassyöpä tuottaa edelleen huomattavia haasteita potilashoidolle, tutkijoille ja yhteiskunnille. Eturauhassyövän rekurrentin muodon parantavan hoidon puuttuessa sairaus säilynee intensiivisen tutkimuksen kohteena myös tulevaisuudessa.

Vaikka lukuisia yksittäisiä komponentteja tai reaktioteitä käsitteleviä tutkimuksia on julkaistu, todellisia yrityksiä eturauhassyövän globaalin säätelytekijäprofiilin muodostamiseksi ei ole tehty koko genomin laajuisella skaalalla. Tässä työssä pyritään muodostamaan pohja tämän kunnianhimoisen haasteen toteuttamiseksi hyödyntäen laskennallisia menetelmiä transkriptiofaktoreiden säätelykohtien ennustamisessa ja säätelyverkkojen johtamisessa. DNase I hypersensitiivisyysinformaation ohjaama transkriptiofaktoreiden sijaintiriippuvien sitoutumismatriiseiden avulla toteutettu skannaus suoritettiin usealla eri tavalla tuottaen katalogi mahdollisista säätely-yhteyksistä.

Ennustetut säätely-yhteydet arvioitiin huolellisesti ja useita kehitelmiä alkuperäisestä menetelmästä esitettiin muodostaen vahvan ohjenuoran tulevien laskennallista säätelytekijäprofiilinrakennusta hyödyntävien sovellusten käytettäväksi. Lisäksi jatkosovelluksena geenieksprессiodataa sovellettiin erottamaan säätelyverkko, joka on läsnä ainutlaatuisesti vain LNCaP-solulinjan soluissa. Johtopäätöksenä genominlaajuiset laskennallista säätelytekijäverkkoa rakentavat menetelmät esitetään toimivana mutta haasteellisena lähestymistapana, jolla kuitenkin on suunnattomia mahdollisuuksia sen sovellusen vuoksi solutyyppikohtaisten ja sairauskohtaisten säätelyverkkojen profilointityökaluna.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Biotechnology
VILLE KYTÖLÄ : Computational prediction of prostate cancer specific gene regulatory network
Master of Science Thesis, 70 pages, 0 Appendix pages
June 2014
Major: Computational Systems Biology
Examiner: Olli Yli-Harja
Keywords: Gene regulatory network, Computational prediction, Prostate Cancer

Prostate cancer is a highly common disease in the Western countries with its prevalence growing every year with the standard of living and life expectancy. Despite extensive studies prostate cancer presents still considerable challenges to patient care, scientists and societies. As still no curing treatment exists for the recurrent form of the disease prostate cancer is likely to remain an intensive subject for study also in the future.

Even though numerous studies exists that try to characterize individual components and pathways of regulation, no real attempts have been made to characterize the regulation profile of prostate cancer in a genome-wide level. This thesis forms a basis for this ambitious feat by using computational methods in transcription factor mediated regulation prediction and network inference. DNase I hypersensitivity guided genome scanning with transcription factor specific position weight matrices is performed in several manners, producing a catalogue of putative regulation connections in LNCaP cells.

The predicted connections are subjected to thorough evaluation and several applications of the first method are presented, forming a strong set of basic guidelines for computational prediction based regulome assembly for future studies. In addition, as a further application gene expression data is applied to selectively extract a subnetwork of regulation present exclusively in the studied LNCaP cells. As a conclusion the genome-wide computational regulome assembly is presented as a functional but challenged approach with enormous potential due to its possible applications as cell type or disease specific network profiling tool.

PREFACE

This thesis was executed as a joint project with the Computational Biology group at University of Tampere and Computational Systems Biology group at Tampere University of Technology. The thesis was supervised by professor Matti Nykter, head of Computational Biology group. All practical work was done in Tampere at the Computational Biology group's premises.

I would like to express my gratitude to professor Nykter for the opportunity to take part to the challenging world of cancer research, and for excellent guidance and advice during the project. I thank my co-workers at the Computational Biology group for support and intriguing conversations. I also wish to thank professor Olli Yli-Harja for helping me in practical matters during the thesis work and for examining the thesis. Finally, I would like to express my gratitude to my friends and family who have put up with me during this challenging work.

Tampere, May 2014

Ville Kytölä

CONTENTS

TERMS AND DEFINITIONS

ChIP-seq	Chromatin ImmunoPrecipitation sequencing, a technology to map the binding sites of DNA binding proteins
AR	Androgen receptor
ARE	Androgen Response Element
BPH	Benign Prostate Hyperplasia
cDNA	Complementary DNA, synthetized from mRNA
CRPC	Castration Resistant Prostate Cancer
DHT	Dihydrotestosterone
DNA	Deoxyribonucleic acid
DNase I -seq	Sequencing technology for DNase I enzyme hypersensitive regions to locate areas of active regulation in genome
GEO	Gene Expression Omnibus
GnRH	Gonadotropin-releasing hormone
IGV	Integrative Genomics Viewer
LNCaP	Androgen positive prostate cancer specific cell line
mRNA	Messenger RNA, RNA molecule carrying transcribed genomic information
NGS	Next Generation Sequencing
PCa	Prostate Cancer
PCR	Polymerase Chain Reaction
PSWM	Position Specific Weight Matrix, identical to PWM
PWM	Position Weight Matrix, identical to PSWM
RNA	Ribonucleic acid
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor

1. INTRODUCTION

Biology is a seemingly never ending field for research. Only a small fraction of the complicated function of a cell is currently understood. Even though years of research have uncovered a plethora of detailed information of the function of a cell the underlying factors that control it such as mechanisms of transcriptional regulation are not at all fully characterized. On a very basic level DNA is the starting point to understanding cell's behavior since it is used to produce the functionality of the cell through proteins. The question how the genotype of an individual manifests as a phenotype is an enormously interesting and relevant in order to understand the delicate machinery that keeps the cell in homeostasis - and what disturbs the balance. Thus cellular behaviour arising from the genotype is inseparably related to human diseases such as hereditary diseases, immunological diseases and cancer.

Due to the formidable complexity of the cell traditional experimental methods of molecular biology have started to give room to a new type of research. As technology advances huge masses of measurement data from inside the cell have become available - a fact which is currently changing the entire field of life sciences permanently. With the high-throughput technologies which produce these vast data collections in relatively short time the researchers are able to see the cell's genomic and transcriptomic features at a whole new level. However, the new possibilities bring also new challenges: information has to be extracted from the mass of measurement data. This is possible using sophisticated mixture of computational methods based on mathematics and signal processing encoded into computer programs.

The fields of computational biology, systems biology and bioinformatics make use of the data through computational methods aiming to increase our understanding of complex biological phenomena and to improve the quality of life we live. These approaches have already proven to be invaluable in fighting complicated and wearing diseases such as cancer. Cancer as one of the most common causes of death in developed countries provides a significant challenge to researchers and medical care due to its difficult treatment and relatively high mortality. Cancer also presents a growing problem with already over 14 million yearly cancer cases in the world and estimates of doubled cancer burden during the next two decades [1]. These facts make cancer a serious threat to individuals and to society, and make cancer research one of the most important fields of research.

In a cancer cell the normal control mechanisms of at least cell division and death have become disturbed. However, as the cell possesses numerous ways to control these functions it is very difficult to determine the reason causing the observed behavior. One possible cause is that something has gone awry in the cell's gene expression control system. The regulation of gene expression is mostly performed by specific molecules called transcription factors (TFs) but also many other factors take part in the process. In a normal cell the regulation machinery makes sure that each protein is produced at a right time, rate and quantity. Since the machinery is networked very tightly together many regulation connections affect and complement each other which makes it is very challenging to find a malfunctioning part using traditional biological experiments.

To respond to the challenge of understanding ever more intricate cellular processes, new generation of measurement technologies has emerged and matured during the last decade to a level where genome-wide data generation is routine-like and relatively inexpensive experiment, producing vast amounts of measurement data. Genome and transcriptome profiling and gene regulatory elements are basic targets for such experiments and also advanced applications grow more common each year. The data explosion in biology desperately craves also better computation power and analysis tools, which are being actively developed by an international community of researchers. It has become clear that there is no turning back: as biology is becoming more and more a data-driven science it depends heavily on computational analysis and information technology. While this progress presents major challenges in data management, control and interpretation, the new techniques both biological and computational provide enormous potential for better science, fundamentally better understanding of biology and ultimately, better quality of life.

To make use of this potential this thesis aims to give new insight into gene regulation in the context of cancer. The presented work makes use of state-of-the-art systems biology methods to construct a network of gene regulation in prostate cancer cells. The network is created by calculating predictions of TF binding sites to the genome and inferring regulation connections from these sites according to proximity of putative regulated elements. These binding site predictions are then filtered for inaccessible sites of the DNA using several different approached, and finally the regulator-target pairs are connected to one large global network of prostate cancer regulation. This novel network could have the potential to pinpoint new connections and subnetworks of regulation relevant to origin and progression of prostate cancer. In addition, the developed method is not application specific and on the contrary can be widely applied to various different network modeling tasks providing an excellent start point for further biomedical studies.

2. THEORETICAL BACKGROUND

2.1 Regulation of Transcription

Transcriptional regulation is a multistage process where a horde of different molecules interacts together directly and indirectly to produce a correct regulation output in each varying situation. However, in order to understand the mechanisms of regulation DNA structure, packaging and modifications must first be understood.

2.1.1 DNA packaging

DNA does not drift idly and alone in the nucleus of a cell. Rather, normally it is packaged tightly around protein complexes consisting from histone proteins which are as part of the histone protein family one of the most common proteins in eukaryotic cells. The histone proteins form octameric histone complexes that take a disc-like shape illustrated in Figure 2.1. As a histone complex rounds up DNA together they form a larger complex called nucleosome which forms a fundamental unit of DNA packaging. The DNA strand is coiled around the histone complex 1.7 times in each nucleosome which corresponds to 147 base pairs. The adjacent nucleosomes are connected by a string of linker DNA that allows flexibility in packaging the structure. Along with other DNA binding chromosomal proteins the forming structure is referred to as chromatin. Chromatin can be packaged even further to different degree according to the cells needs through formation of complicated coiled-coil structures.

A loosely packaged form of the chromatin is so called 'beads-on-a-string' structure where the linker DNA is fully stretched out. This is illustrated on the right in Figure 2.1. A much tighter packaging is achieved when the nucleosome complexes curl up and form a tight meshwork structure. This packaging of the chromatin increases the density of DNA remarkably and enables the small nucleus to contain huge macromolecules such as chromosomes: the most packaged form of DNA when a whole chromosome condenses, for example, in cell division, packs the DNA 10,000 times shorter than its fully extended form.

In addition to saving space, through packaging some areas of the chromatin can be condensed to an very dense, inaccessible bundles called heterochromatin. The DNA in heterochromatin is purposefully blocked. The blocking function of hete-

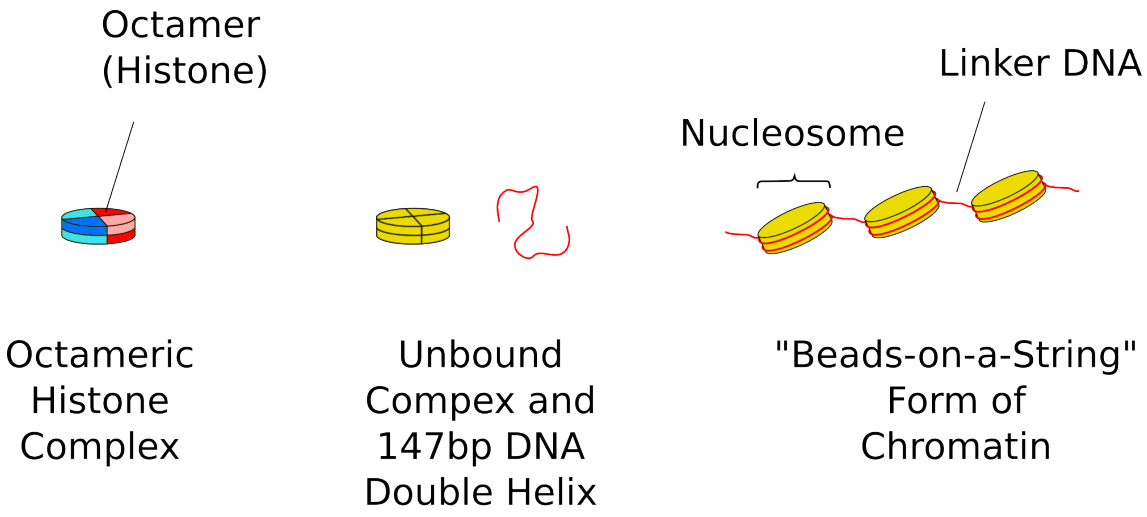

Figure 2.1: Histones in a complex, nucleosomes and "beads-on-a-string" conformation [4]

rochromatin largely assists cell differentiation when unwanted areas of the genome are effectively repressed or silenced: for instance, a liver cell has many functions that a neuron does not, and it is also of vital importance for the liver cell to suppress expression of all neuronal features coding genes in order to function like a proper cell of the liver. Heterochromatin is a cost-effective way to implement this functionality that helps the cells in differentiation. [2]

Chromatin can be rapidly modified by special proteins called chromatin-remodeling complexes which are ATP-driven machines functioning to condense or unpack chromatin. Chromatin is also modified by proteins performing histone modifications such as acetylations, methylations and phosphorylations. The modifications take place in tails of the histones and have a multitude of functions, many still unknown. A vast variety of different modifications targeting different subgroups of the histone tail exist and the effects they mediate are known to be very important to the cell: the modifications can change the electrostatic properties of nucleosomes resulting in packing or unpacking of chromatin, they affect the binding of chromatin factors and they may also promote or repress processes like gene expression [3] In addition, histone modifications have a crucial role in epigenetic heritance where non-DNA information is passed to progenitor cells. [2]

2.1.2 Gene Regulation

A decade ago almost beyond our reach, the uncovering of an organisms genome is nowadays a fairly simple task thanks to modern high-throughput measurement technologies. However, even when provided with the information of an organisms

detailed DNA structure one does not in practice achieve any closer to understanding how such complex organisms as humans are built: even though all somatic cells in our body are identical from their genomic contents they form a huge variety of different tissues. In order to form a multicellular organism the different cells must control the production of their proteins in highly controlled and specialized manner: to be able to ensure the exact function of the cell the expression of a gene and eventually production of the corresponding protein cannot be random processes. Indeed, regulation of gene expression takes place in every phase along the process starting from DNA and ending in protein synthesis, function and ultimately degradation. This complicated and fine-tuned machinery allows organisms to form complex structures like tissues and organs, and enable them to perform the tasks that belong to what we consider life. [4]

Eukaryotic gene expression is a subject to several different control phases visualized in Figure 2.2. One important regulation step occurs already when a gene is described into RNA. Another controlled process that effects the expression and thus the behavior of the cell is post-transcriptional modification of messenger RNA (mRNA) which is governed by delicate protein machinery that performs splicing, capping and transport tagging functions[5]. mRNA transport and localization are also examples of controlled processes: RNA molecules need to be delivered to right ribosome complexes to enable correct protein production. Before translation can occur processes that force RNA to be degraded can take place and thus affect the expression of the gene. Also mRNA translation of the surviving mRNAs especially at its initiation stage is strongly controlled by a horde of pro-translational factors that with the ribosome subunits form a translation initiation complex. Finally, after translation the activity of the produced protein is also a subject to regulation through different post-translational modification processes. Finally, after the immense task of producing the functional protein its degradation frankly is another important regulatory process. Efficient degradation of proteins is vitally important since many cellular processes need to react quickly to changes in the environment and cell's internal state. On the other hand a protein that is consistently needed in the cell can be maintained at sufficient levels with reduced degradation rate that allows protein pileup and thus desired functionality. [4]

The signal that initiates different regulation control mechanisms may rise from various locations or factors. Based on its origin the regulation of expression can also be divided to extrinsic and intrinsic control. Many extracellular signals such as secreted signaling proteins, other small molecules and physiological factors like temperature, pH and oxygen concentration can affect the expression of a gene. All these signals affecting the cell induce relatively rapid but often temporary or at least reversible changes in gene expression. Through this nature of extrinsic control the

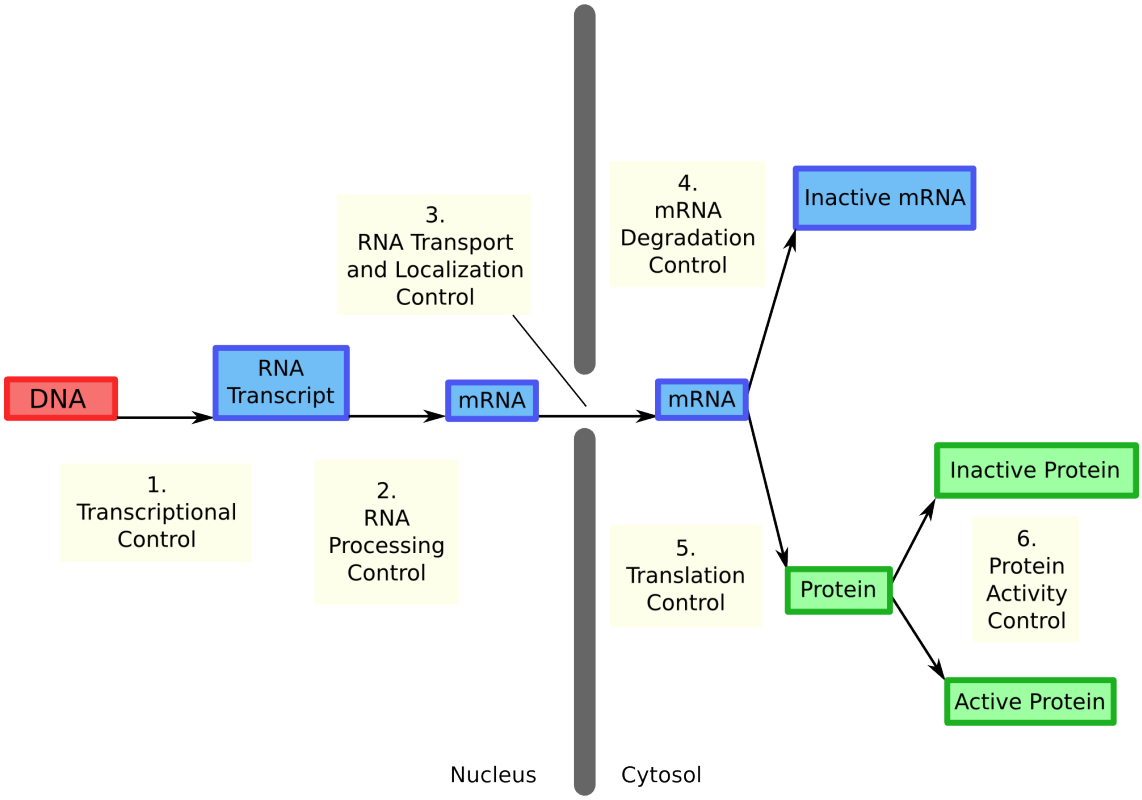

Figure 2.2: Steps of regulation during gene expression to proteins [4]

reaction mechanisms of the cell can have very significant effects on the cell itself, its neighbors or even the entire organism through changed levels of gene expression resulting in, for instance, hormone production and release. Hormone secretion is one way of passing information in the organism to guide and coordinate the function of other cells to perform vital functions like maintaining homeostasis. Thus with extrinsic gene expression regulation the cells are able to communicate and adjust their behavior according to changing needs. [6]

While it is evident with certain exceptions that the genome of a cell does not change when a cell differentiates, the cell needs a way to permanently change its patterns of gene expression to become truly different than its progenitor. In addition to TF induced control this can be achieved also through intrinsic control of gene expression. This means in practice modification of DNA mainly through methylation and modification of chromatin most commonly methylation or acetylation with histone modifications. These changes have the potential to alter DNA structure is such a way that access to DNA and thus to particular genes may be gained or blocked. This form of so called epigenetic or non-genomic regulation is very important in cell differentiation. As the use in differentiation hints these modifications are typically more stable than regulation conditions induced by TF binding. However, they are not fixed and can be changed which is a common event for instance in some cancer

types. [6]

2.1.3 Transcription Factors and Regulation of Transcription

One of the most important steps of gene expression regulation occurs already at DNA level: the transcription of DNA into RNA poses a major control point in the whole gene expression process. Regulation of transcription is not only efficient but also versatile place for expression control. The efficiency is mostly due to the resources saved by the cell: if the expression of a gene needs to be lowered or stopped the most energy and building materials are saved when the cellular workflow leading ultimately to protein synthesis is arrested at the first possible stage. On the other hand the complexity of transcriptional control allows a wide-ranging collection of different possible states of expression in a cell and enables also fast and powerful response to stimulus. These features of transcriptional regulation are achieved through use of transcription factors (TFs) to control which genes are transcribed and in which amounts. [4]

Transcription factors are proteins or protein complexes that share the common feature of possessing a DNA binding domain. The domains of different factors are sequence specific and thus enable the TFs to bind different areas of DNA called TF response elements. Unlike, for instance, the restriction enzymes that recognize and cut DNA at their specific target site and that site only, a TF binds a much wider range of response elements: the elements that match the domain of the TF more accurately bind with higher affinity while more mismatches containing response elements result into weaker binding. This mechanism allows much stronger control of transcription since it allows not only TF specific regulation but also binding affinity based regulation with each TF. This results into a complex mixture of regulating factors that bind response elements with different speed and stability allowing the gene expression to be adjusted with very fine tuned steps if needed. [7]

TFs were traditionally thought to bind to gene promoter and that way regulate its expression by either promoting the formation of transcription pre-initiation complex or repressing it. After detailed studies the promoter area of a gene is usually divided to core promoter region which corresponds to the binding site of the RNA polymerase II complex, and to promoter-proximal region which contains TF binding regions some 1000 nucleotides upstream of the core promoter. In addition, already for some 30 years it has been known that TFs regulate the expression of genes also through other sites in the DNA. These sites can be located even hundreds of thousands of base pairs upstream of the promoter sequence and they take part in regulation through long loops in the DNA which bring the bound TFs to proximity of the promoter. These far-upstream elements are divided into enhancers, silencers, insulators and tethering elements. Silencers are sequences where repressor TFs bind to hinder

the transcription process through interactions with the promoter binding elements or enhancers. Insulators are chromatin elements that bar access from regulating elements to certain area, thus insulating the influence of different signals, whether positive or negative. Tethering elements are usually located near to promoter region of the cell and help to guide the looping enhancer elements to right gene's promoter. [8]

The most important far-upstream regulatory elements are the enhancers which contribute positively to the transcription of genes. Enhancers and their bound TFs have a key role in transcription initiation. Enhancer sites are short, some hundred nucleotide long sequences that function as docking areas for specific transcription factors. One enhancer region often contains not one but a cluster of response elements of different TFs. Interactions between these enhancer TFs, with other far-upstream elements and with the promoter elements results in a combinatorial effect that has the ability to produce accurate expression patterns which are important, for instance, in many developmental processes. [8]

In addition to the spatial binding patterns of TFs, the regulation has also a temporal nature. It has been shown that at least in some cellular processes the binding of TFs to enhancer regions is time dependent even though the levels of the TF remain constant the whole time. This suggests that the time-dependent binding of TFs is not only controlled by the expression level of the TFs but also by more complicated mechanisms. The temporal binding is thought to be regulated by binding affinity to the sites response element and also number of response elements which can make the binding more sensitive to small changes in TF concentration. [8]

2.1.4 Gene Regulatory Networks

The previous discussion of regulation of one gene already reveals the nature of more general gene regulation: the gene is controlled by TFs which in turn are sensitive to other TFs and regulation mechanisms. In other words, cell-wide transcriptional control is far from being an independent process. Rather, the regulation connections typically form a network that controls the expression of genes in a complicated manner. Typically the TF control mechanisms for functional entities which can be quickly activated: for example, one signal can activate several TFs to regulate their targets in a differential way. This means the regulation enables the mechanism of cellular responses which result in cascade-like changes in expressions of multiple genes. This type of regulation allows the cell to react to changing conditions or needs. Ultimately, together the different TF-target connections form a cell-wide regulation network with distinct global and local topology. [9]

Studies indicate the global topology of the gene regulatory network shows unex-

pected connection patterns. Instead of normally distributed number of regulation connections between elements it appears that nodes with high number of regulation connections are relatively common. These so called hubs are often more conserved than other genes and also have a role in more vital functions of the cell. The gene regulatory networks are often classified as critical because of their specific topology and resulting function. Emergent phenomena are frequently observed in critical networks. In the context of gene regulatory networks this means the regulome's ability to form complicated and in long term unpredictable behavior. [10; 11]

Also the local structures of gene regulatory networks are of great importance to the cell. Local networks of a few genes can form numerous network structures or motifs that can implement many surprisingly complicated functions. For instance, a structure resembling a toggle-switch exists and possesses an important role in several cell fate determining processes. In the toggle-switch structure two genes mutually repress each other. Often the initial low expression of both genes can be turned into domination of the other TF by some external stimulus. The TF that gains the dominance over its switch pair leads to activation other TFs possibly resulting in a lineage fate choice and differentiation [11]. Other examples of interesting local motifs are feedforward loops and negative autoregulation. Feedforward loops have several different topologies and are capable of implementing regulatory 'logical gates' that produce AND and OR type regulation functions. Feedforward loops can also produce pulse-like dynamics and thus periodical gene expression. This is of special importance to many cellular functions related to detecting an external signal and provide an energy saving solution since the detector gene is produced more rarely in periodic pulses. In addition, negative autoregulation, although being an extremely simple one gene motif, has interesting features as it is able to tune the expression of a gene to a steady level by a concentration dependent threshold of self-repression. These common local motifs are illustrated in Figure 2.3.

As hinted by the function of toggle-switches, gene regulatory networks have a vital role also in early development of an organism as they control the spatial localization of different cellular functions in an embryo. Regulatory networks also govern processes like pluripotent stem cell differentiation and eventually formation of different tissues [11]. Because all organism's cells originate from a common ancestor, the fertilized egg, it can be stated that the underlying global transcriptional regulatory network of all possible regulation connections in a cell is somewhat static. This arises from the locations of the response elements to different TFs which remain unchanged in all unmutated somatic cells of an individual. The actual functional regulation network of a given cell type is a specific instance of this global putative connection network and it defines the detailed cell type and state characteristics. This is completed through competitive binding to overlapping response elements

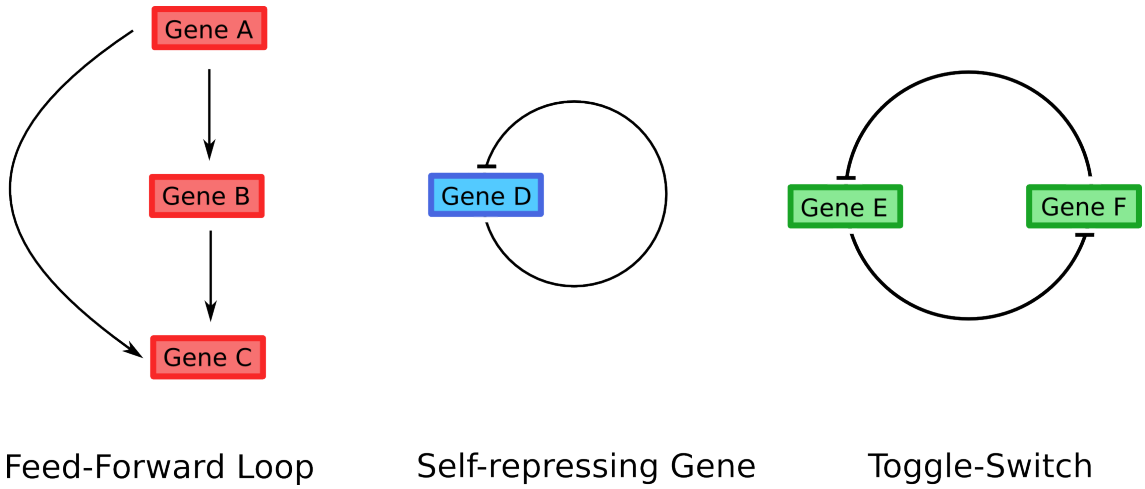

Figure 2.3: Common local network motifs: one type of feed-forward loop, self-repressing gene and a toggle-switch

resulting in differential regulation [12], and by physical changes in the DNA structure and packaging like formation of heterochromatin to permanently silence certain genes leading to unwanted progression lineages [3]

Often a specific part of the global network is a subject of interest for researchers. For instance, numerous pathways have been annotated to databases like KEGG (Kyoto Encyclopedia of Genes and Genomes) and several others, containing parts of gene expression networks in those annotations. Also networks spanning from single regulating factors such as the subnetwork of androgen receptor (AR) are typically a target for studies. For example, the mentioned AR subnetwork has been intensively studied and its characteristics have been presented in many publications due to its crucial role in prostate cancer [13]. The network models can provide help in characterizing dynamic behavior of a system as interactions can be perceived from the network connections.

2.2 Pathology of Cancer

Cancer is a serious, often life threatening ailment to the patient and has also a significant societal impact due to its costly treatment and growing incidence due to improving standard of living and prolonging life-expectancy. In this chapter a short introduction to cancer as a disease is given. Furthermore, prostate cancer is discussed in more detail as it is of special interest in the context of this thesis. Lastly, cell line models and the validity of cell lines as models of a disease is discussed.

2.2.1 Cancer as a Disease

In the world of the cells, death is a natural part of life. The cells of an organism form a fine tuned functional entity based on collaboration and mutual benefit: the cells

share resources and keep each other alive while performing their given tasks which in turn enables the continuity of their existence. One could say, the cells work together for greater good, which in this case is the independent life of the individual they form. In this collaborative effort should a cell come across some misfortune leading dealing irreparable damage to its function the cell as its normal response seemingly selflessly gives its life in order to avoid causing further injury to its neighbors and, ultimately, to the whole organism. This controlled suicide mechanism of the cells called apoptosis enables the continuous life of an individual despite environmental hazards and stochastic processes of the physical world.

Mutations are a driving force of evolution and ensure the development and adaptivity of every species. However, mutations in the genome of an individual cell occur randomly due to the purposefully imperfect DNA replication and repair machinery the cell possesses. Given the human genomes immense size compared to gene coding regions it may often be that an occurring mutation has no effect on the protein manufacturing of the cell and induces no changes in the cells function. However, it has been shown that there is a link between mutagenesis and carcinogenesis (the generation of cancer). That is, should the mutation occur in some protein coding sequence or, for example, in a regulatory element, the effects are rarely positive. Rather, the mutation may result in loss of function of a protein domain, result in misfolding of cellular structures, illogical regulation patterns or any combination of these. A normal cell however has the ability to deal with these misfortunes through apoptosis, and make sure the harmful mutation is not passed along to the cells progenitors. [14]

One mutation does not cause a cancer. This is quite evident just by inspecting the average mutation rate in the human genome. Due to replication and error correction machinery's imperfect accuracy it has been estimated that approximately 10^{-6} mutations occur per gene with every DNA replication. As an average human experiences about 10^{16} cell divisions in their lifetime it seems that the number of occurring mutations during a persons life is tremendous. From this inspection it seems clear that an individual mutation cannot be the cause of a cancer - we simply would not be able to exist if every one of these frequent mutations was lethal. Instead, often a process of tumor progression is considered: initially a harmless change in the cell eventually transform the cell into an actual cancer. [14]

In addition to genetic mutations the piling up in the cell's DNA also epigenetic patterns have been shown to play a role in tumor progression. A normal cell trusts epigeneticly inheriting histone modifications to enforce similar structure of heterochromatin and DNA methylation in its progeny to ensure that critical genes stay silenced. In similar manner, genes controlling tumor suppression may end up being silenced by heterochromatin during tumor progression. These epigenetic modifica-

tion patterns are transmitted to the daughter cells as surely as the genetic mutations and may have significant impact in the development of the tumor. [14]

Even though cancer promoting mutations are generally regarded as highly negative and harmful events, from the perspective of the mutated cell the effect is quite opposite. The mutation in practice means the cell has gained an advantage over its peers, giving it seemingly unnatural powers such as superior nutrition intake, growth rate or immortality. This can be pictured as a process of microevolution where mutations give the cell an evolutionary advantage over its competitors. With the emerged edge in competition the cell quickly starts to dominate its surroundings. This is generally regarded as the creation of a benign tumor. Should the cell obtain a fitting series of mutations resulting in more malfunctioning cellular controls it is possible that the tumor turns malignant and starts to invade adjacent tissue. At this stage the cell's internal safeguard mechanisms have profoundly failed due to the stockpiled mutations and the cells cannot contain themselves anymore - a cancer has been born. The evolutionary advantage of the original mutated cell has granted the tumor dominance over its former collaborators but instead of resulting in an advantage for the organism the growing population now threatens the function of the tissue and life of the entire individual. [14]

2.2.2 Prostate Cancer

Prostate cancer presents an ever growing concern to the western world. It is the most common male cancer and second most common cancer related cause of death just after lung cancer, with over 230,000 estimated new cases and almost 30,000 deaths in 2014 in the United States alone. However, even though estimates state that 1 out of 7 men are expected to be diagnosed with prostate cancer during their life only approximately 1 man of 36 will eventually die from the cancer. This is due to prostate cancer being a disease of elderly men with 6 of 10 cases being diagnosed in men in their mid-sixties or older. The advanced age of the patients combined with modern treatments mean that many patients will die with the cancer but not from it. An even more common prostatic change than the cancer itself is the so called Benign Prostatic Hyperplasia (BPH) which means benign increased proliferation of the prostate gland cells: even 90% of men in their eighties are estimated to exhibit BPH [15]. While BPH can develop into prostate cancer it is most often harmless and will not affect the patients quality of life. [16]

The male sex characteristics arise mainly from androgen hormones' effects in the body. Testosterone, secreted by the testes is the principal androgen that is transported in the body through blood circulation in albumin and globulin complexes. Testosterone is converted to dihydrotestosterone (DHT), a far more reactive androgen while entering the prostate cells. DHT has a high affinity for the an-

drogen receptor (AR) which is a nuclear receptor protein composed of activating domain, ligand-binding domain and DNA-binding domain formed of two zinc finger motifs. The binding of androgens induces a conformational change in AR resulting in phosporylation of the receptor and release of heat-shock proteins attached to the basal state receptor. The changes enable formation of AR homodimer complexes which are eventually able to interact with Androgen Response Elements (AREs) in the DNA. The binding of the AREs results in changes in gene expression ultimately stimulating the prostate cell proliferation. The principle of AR activation pathway is pictured in Figure 2.4 [17]

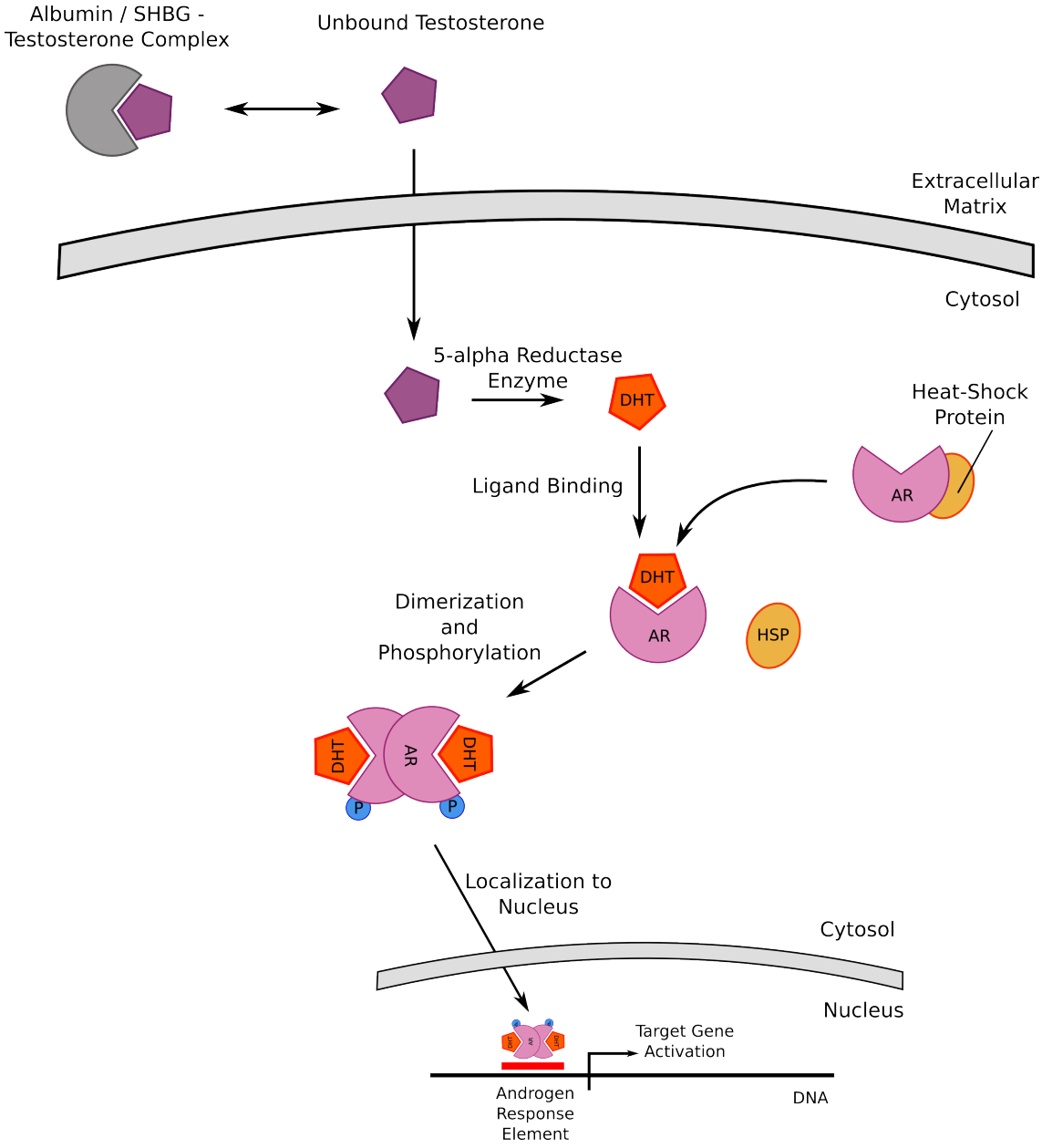

Figure 2.4: Androgen receptor response to androgen stimulation

As mutations take place in the prostate tissue and some cells start to show can-

cerous behavior the cells remain still under strong androgen based growth control. This is why androgen ablation therapy, that is, the artificial lowering of the androgen levels in the body is typically used in the treatment of first stage prostate cancer. The cancer cells respond to the deprivation of growth-inducing androgens by regressing and the advancing disease halts. With the androgen ablation therapy the first stage prostate cancer is treatable and generally yields very good results. The effects of androgen deprivation were first observed more than 40 years ago when Charles Huggins was able to show that orchiectomy, that is, the removal of the testes had the potential to halt advancing prostate cancer. Currently the androgen ablation is achieved by so called chemical castration using gonadotropin-releasing hormone (GnRH) agonists and androgen antagonists instead of the physical castration treatment.[17]

The mere androgen deprivation is not enough, however. Without additional treatment the cancer eventually becomes androgen independent or and symptoms of the disease return. Time span of this recurrence varies from several months to several years and it is observed in most patients whose treatment has failed in decimating the cancer tissue in the first stage of the cancer by androgen deprivation and additional methods such as radiation therapy and surgical operations. The recurring androgen independent or so called castration resistant prostate cancer (CRPC) is a far more serious form of the disease. A metastasized CRPC remains still an unconquered challenge to medicine since currently only palliative treatments exist - the disease unstoppably leads to the patient's death.

Reasons behind the recurrence are still not thoroughly known. It has been shown that despite the androgen deprivation therapy the CRPC cells contain small but detectable levels of androgens which can be enough to induce AR mediated cell proliferation and recurrence of the cancer [18]. In addition, the recurrent form of the cancer seems to show high levels of androgen and elevated levels of AR coactivators putting emphasis to androgen hypersensitivity theory where even tiny levels of androgens are enough to launch the AR mediated response [19]

Prostate cancer has been studied intensively and various mutations and chromosomal rearrangements such as gene fusions have been reported as contributing or even driving factors in prostate cancer development. One factor deserving special attention due to its crucial role in the progression of the cancer, however, is the androgen receptor - indeed, several scientific publications have studied the role of AR and its regulation targets in various conditions. However, a broader mapping of gene regulation in prostate cancer is another issue, and to our knowledge no previous work on global prostate cancer specific regulation profiling has been published. This fact makes prostate cancer an interesting subject for this study and adds to the potential of eventually discovering novel pathways of differential regulation in

prostate cancer.

2.2.3 Cell Line Models

There are many problems with cell cultures obtained from human tissue. First of all, since the environment in the culture is drastically different to that in the tissue due to loss of 3D structure, changed nutrient flows and attachment to surrounding tissue, the cells are very hard to grow. In addition, even growing cells eventually stop dividing as shortening of telomeres or other issues activate cell cycle arrest, meaning in practice that the culture becomes superfluous. On the other hand, the cost of obtaining tissue samples measured in time, money and patient comfort are considerable. An answer to these issues is provided by using special cell line models in cultures and in the following experiments.

Cell line models are more easily grown variants of their anatomical correspondents that try to serve as an accurate model of the original tissue. Cell line cells are capable of indefinite number of divisions making them in this sense superior to culturing tissue samples. Especially in cancer research cell line models have been used largely. The reason for this in addition to the challenges in retrieving the samples that cell lines are rather easy to create from cancer cells since they already possess many traits that force them to grow rapidly and refuse to die. Since cell line cells can be rapidly grown and also can be stored at extreme temperatures without loss of function after melting they provide an easy and cost-effective way to study different tissue type characteristics and before all, different diseases. However, since the cell line shows radically different behavior in proliferation than its original parent cells, many other changes in the cells' function may have occurred as well. It has to be kept in mind that results produced in cell lines must always be critically reviewed and validated with the original tissue. [14]

A myriad of various cancer cell line models exists, each with its characteristic features. Only in prostate cancer over 200 cell lines and sub lines have been used [20]. Only a handful of these, however, are in active widespread use: with this restriction the number of prostate cancer specific cell lines reduces to little over 20 lines. One of the commonly used lines is called the LNCaP cell line. It has been derived from prostate cancer metastasis in lymph node of a caucasian male patient. LNCaP cell line exhibits androgen sensitive behavior which makes it an important model for prostate cancer studies. [21]

2.3 Next Generation Sequencing

The Sanger sequencing method developed in 1977 by Frederick Sanger long remained the most commonly used sequencing technology [22]. Enormous achievements like

finishing the Human Genome Project and thus sequencing the first whole human genome was completed using the Sanger sequencing. However, due to cost and extensive amount of work and time required for the Sanger sequencing new methods were urgently called for during the last decade [23]. As an answer to the need next generation sequencing technologies (NGS) were developed. This naming convention refers to the automated Sanger sequencing method as the 'first generation' making the modern methods the 'next generation' of sequencing [24]. Here we discuss the principles of these methods to comprehend the techniques that were used to generate the datasets used in this thesis.

2.3.1 General Principles

The next generation sequencing techniques are a heterogeneous group of automated sequencing methods capable of so called high-throughput data production. This means they have the ability to produce very large amounts of sequenced nucleotide fragments in relatively short time. Many different commercial manufacturers including Illumina / Solexa, Roche / 454 and Helicos Biosciences to name a few, have developed their own version of the sequencing technology. A modern high-throughput sequencer is a table-top machine that can sequence the whole human genome in a few days or even faster, producing tens of gigabytes of data. Most common applications of NGS experiments are de novo genome assembly, identification of Single Nucleotide Polymorphisms (SNPs) and other variability in genomic structure, transcriptome profiling (RNA-seq), mapping transcription factor binding cites (ChIP-seq) and profiling the structure of the chromatin for epigenetic markers [25]

Next we consider the technology behind the common sequencing technologies. The process of next generation sequencing can roughly be divided into steps of template preparation, sequencing and imaging and data analysis. The actual implementations of these steps, however, vary greatly between different commercial manufacturers due to greatly differing technologies. These methodological differences result in variance between sequencing quality, output and price. [24]

The templates used in sequencing are recombinant DNA fragments that consist of a known and unknown region. The known region is a specific sequence capable of binding a primer sequence. The unknown part of the template, called also the target sequence is the actual portion of interest, which will get sequenced in the experiment. Typically, the sample DNA is fragmented to small pieces and attached to known adaptors in the template preparation phase. These template sequences are then attached to a surface immobilizing them and thus allowing massively parallel sequencing of these templates. In many cases the templates are clonally amplified since the majority of imaging technologies cannot detect signals using single template

resolution. One approach in amplification is to use primer covered beads to which a single template first binds. The template is replicated using PCR repeatedly until finally the bead is covered with clones of the same template. Depending on the technology, the beads are then either attached to a glass surface with chemical cross-linking (Life / APG) or divided into specific wells (Roche / 454) to hold them in place during the sequencing. Another approach is to use so called solid-phase amplification utilized by Illumina/Solexa systems. Here a 'lawn' of forward and reverse primers is created on a solid surface and the templates are attached randomly to the surface. The templates then bend and form primer-template pairs in a process called bridge-amplification, where the initially one-stranded bridges are amplified using PCR to eventually create clusters of similar templates on the surface [26]. In addition to these clonal techniques, also single-molecule template utilizing technologies exist but these are not considered here. [24]

In the sequencing and imaging phase the clonally amplified templates are subjected to different methods where fluorescent nucleotides are allowed to bind the templates which are then excited by laser and the emitted light is detected by a imaging device. The method used by Illumina systems is cyclic reversible termination where nucleotides are attached to the template strands in cycles. The first phase of the cycle is attachment of a single terminating nucleotide to each template strand by DNA polymerase and the remaining free nucleotides are washed away. The terminating property of the attached nucleotide is of crucial importance since it prevents dephasing of the different template clusters.The attached terminating nucleotides have also a phosphorylating group attached to then which is then excited and the whole surface is imaged to determine the attached bases to each position. The attached nucleotides are then cleaved for their terminating and phosphorylating parts and these cleaved remnants are washed away before the beginning of the next cycle. This cyclic process allows rapid and automated sequencing which is able to produce enormous number of reads in relatively short time. [24]

Another approach used by Life / APG's SOLiD platform is so called sequencing by ligation where DNA ligase is used in cycles. The cycle starts by letting specific fluorescent probes hybridize to the template strands. The probe is attached as reverse complement to the template by a DNA ligase and after the unbound probes have been washed away the platform is excited by laser and resulting emitting light is detected to determine what kind of probe was attached to each position. The SOLiD platform uses a two-base-encoding where each two-base unit's sequence is determined by a specific combination of consecutive probe binding. After excitement parts of the probes are cleaved away and the cycle starts from the start. Due to the special two-base encoding the first round of cycles cannot be used to determine nucleotides at all positions of the template. Instead, a second round is followed

started with a single nucleotide shift. The probe-attachment cycles are then run again to the whole template's length. The process is repeated altogether five times resulting in five strands of linear color sequences. By combining the information of these five rounds the actual sequence can now be determined.

A third method called pyrosequencing used by Roche / 454 incorporates single nucleotide additions where DNA synthesis is halted not by terminating nucleotides but by DNA polymerase manipulation. The polymerase is able to synthesize one nucleotide to the template and then it pauses, after which the system measures light emitted by a series of chemical reactions operating on released inorganic pyrophosphate. The light peaks can be detected in parallel by a imaging device and the cycle can be re-initiated by activating the DNA polymerases to add the next nucleotide to templates. The light peaks are stored and represented as a sequential collection called a flowgram from which the correct nucleotide order can be read.

The last necessary step in the NGS sequencing experiment is genome alignment or assembly. However, methods related to these topics are covered in more detail later with other high-throughput data analysis steps. We continue by considering applications of the next generation sequencing. Various modifications of the basic method exist: in addition to DNA sequencing a commonly used method to profile accurately the transcriptome of the sample is to perform so called RNA-sequencing which is in practice done by reverse transcription to cDNA, which is then sequenced. Many other applications exist also, and in the following sections we introduce shortly two applied sequencing technologies: DNase I sequencing and ChIP sequencing which are constructed based on the basic DNA sequencing methodology but which use specialized ways to obtain the sequenced material.

2.3.2 DNase I sequencing

DNase I sequencing (DNase-seq) is a relatively new technology published in 2010. Although already a successive technology has emerged [28], DNase I sequencing remains a well functional tool for high-resolution mapping of active elements in the DNA. The method was developed to discover and identify active gene regulatory elements in mammalian genomes by combining DNA ligating and NGS approach in a novel way. [27]

The DNase I sequencing is based on DNase I enzyme which has the ability to splice DNA at specific locations. As a specifically extracted and prepared DNA sample is subjected to this enzyme it performs cuts at locations in the DNA that are not blocked by bound proteins including histone complexes. This yields spliced DNA fragments from all areas not blocked by DNA packing or organizing molecules, or by dense heterochromatin. As the resulting fragments are sequenced and the output is aligned to genome the ability of the experiment to show the areas of open chromatin

is revealed. This information can be used in many applications especially related to regulation of transcription in the cell. [27]

Previous low-throughput approaches utilized Southern blotting to deliver similar information of DNase I hypersensitivity. The number of targeted sites was, however, severely restricted due to lack of resources as the Southern blotting is not designed for large scale experiments. The DNase I sequencing provides a high-resolution alternative capable of genome-wide mapping of DNase I hypersensitivity. It is a relatively straight forward protocol and can be applied to practically any sample type. Challenges for the analysis of the technology's outputs are that there are no ways of determining which elements are interacting in the detected sites. In addition, due to the methods rather young age the analysis methodology is not entirely well-established as no clear method of choice exists. Finally, to use DNase I sequencing a reference genome must exist and be available since the method produces reads only from selected genomic regions which need to be aligned to the reference to produce locations information of hypersensitive sites. [27]

2.3.3 ChIP sequencing

Chromatin ImmunoPrecipitation sequencing (ChIP-seq) is a combination of immunoprecipitation of specific elements of interest and subjecting the genomic sequences bound by these areas to next generation DNA sequencing. The ChIP sequencing method enables mapping the binding of various elements such as transcription factors, modified histones and nucleosomes in a genome wide level. First ChIP-seq experiments were published in 2007 and the technology has been in active use by researchers from since. The ChIP sequencing method exceeds the preceding microarray-based ChIP-chip technology in quality and quantity and is a powerful tool for regulome profiling experiments. [29]

In ChIP sequencing experiment the states of DNA-bound molecules are frozen with a fixation agent, usually formaldehyde. This treatment cross-links proteins to DNA preventing them from detaching during the first phases of the experiment. After protein cross-linking the DNA is spliced with a ligase producing a vast number of DNA fragments, some of which bound by various proteins. Next, immunoprecipitation is performed using an antibody specific to the protein of interest. The antibody treatment is used to extract the interesting proteins from the bulk of DNA fragments. These fragments are isolated and afterwards the cross-linking is reversed. The free DNA fragments are then sequenced yielding a collection of information of the binding sequences of the studied proteins. [30]

Due to its high resolution, lower noise and greater coverage compared to the preceding microarray technology the ChIP-seq has provided significantly better data for mapping epigenetic markers and regulating factors on genome-wide level. The

major challenges of ChIP sequencing lie in the analysis of the measurement data - a subject considered in more detail in the following sections. The only practical restrictions of the method are the existence and availability of a reference genome, since the resulting reads need to be mapped against a reference to produce the binding site information. Another restriction is the availability of efficient antibodies for the protein of interest. Even though for the most common proteins strong and selective antibodies exist, in studies targeting previously unstudied elements the finding of a suitable antibody can be challenging. [29]

2.4 Microarrays

A microarray is a compact device capable of measuring features of a target molecule on microscopic scale by using so called probes. The probes are short oligonucleotide sequences that interact with corresponding sample, and quality and quantity of the sample present can be measured based on a fluorescent property of the probes which can be activated through stimulation with laser. DNA microarrays were initially enabled by the development of nucleic polymer amplification methods such as polymerase chain reaction (PCR) and the technology has since developed to produce a wide range of different arrays for specific profiling purposes. [32]

Microarray technology was greatly responsible for the genomics data explosion in the beginning of the last decade. The ability to map large proportions of the genome, transcriptome and though applications also DNA-protein interactions produced a vast amounts of new information on different levels of cellular activity. Also the integrative use of different 'omics' data was emphasized when connections between DNA-level modifications and for example, gene expression patterns were noticed on genome-wide scale. The basic microarray technology was also found to be very adaptive resulting in applications like copy number analysis and mapping of epigenetic modifications. Although being currently a recessive technology due to advancements in next generation sequencing, microarrays remain a cost-effective and sufficient routine-like approach to many genetics studies. Typical workflows of single-channel and two-channel microarray experiments are illustrated in Figure 2.5. Single channel microarrays measure only one sample at a time using one dye while the two-channel arrays are capable of measuring two samples at the same time using two distinct dyes. [31; 32]

Although several commercial manufacturers make microarrays, here we concentrate on the structure and performance of Illumina Bead Arrays since they were used to produce the microarray datasets used in this thesis. This type of microarray consists of large number of beads placed in small wells usually on the surface of a silicon/plastic slide. Each of the beads is covered with hundreds of thousands of short identical oligonucleotide sequences known as probes. Each probe has a differ-

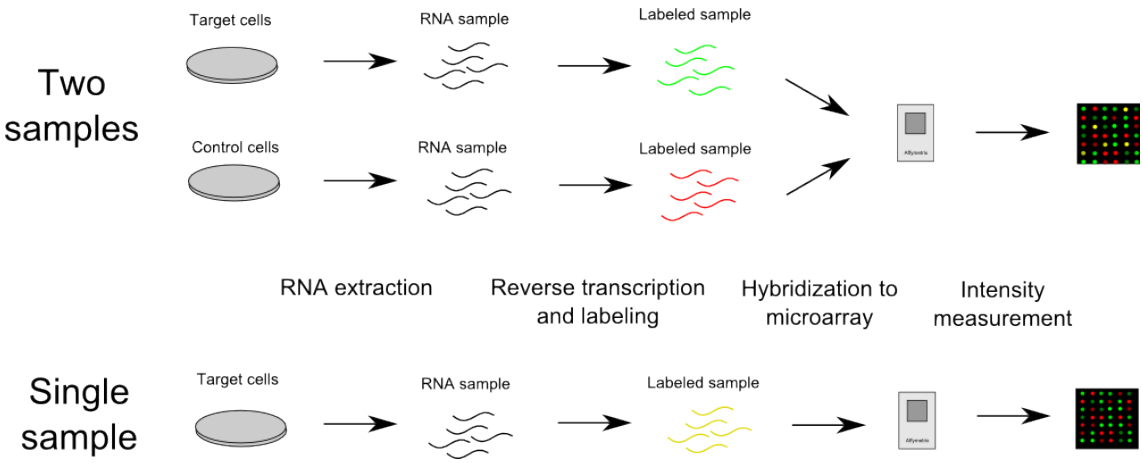

Figure 2.5: Workflow a typical microarray experiment

ent set of oligonucleotides attached to it. In the experiment the beads are washed with prepared sample solution where the DNA of interest has been fragmented to small pieces. The fragments for which a reverse complementary probe exists hybridize to their counterpart while the extra fragments are cleaned away from the array. The hybridized probes are then excited with laser and specific fluorescent groups in the probes activate, emitting light based on the degree of hybridization. Since each bead contains a huge number of probes this allows dynamical measuring of the DNA fragments: small amounts of certain fragment result in smaller emitted light intensities while large amounts produce large intensities. The emitted light is measured using an imaging device which interprets these to numerical data through an automated software. This way, for example, in a gene expression measurement the expression values are obtained indirectly through measured intensities emitted by probe sets corresponding to a certain gene. [33; 14]

2.5 High-Throughput Data Analysis

The vast amount of data generated by NGS technologies is of course no use without efficient analysis tools and pipelines. Analysis of NGS data presents itself as a complex task that depends heavily on the aimed application of the experiment. Here we will consider the analysis steps relevant to this thesis starting from NGS data processing and peak detection and ending in microarray data analysis.

2.5.1 Next Generation Sequencing Data Analysis

Next generation sequencing data analysis is typically very application specific. Common steps for all experiments are usually pre-processing related tasks that often consider issues like removing biases from the sequenced DNA fragments called reads.

In addition to pre-processing many applications like RNA-sequencing require calculation of read counts an their normalization often using modern model based approaches to produce comparable robust gene expression data. However, since the NGS applications utilized in this thesis do not require such normalization methods we ignore them in our discussion and instead present some relevant features of the NGS analyses that were used to produce datasets used in the thesis.

Preprocessing and alignment

The need for data preprocessing arises from errors and unwanted trends in the data that could distort the results of the rest of the analysis. A bias in the measurement data may be caused by various factors such as different measurement conditions, varying amounts of reagents, careless performance of the experiments and biological variance as the cells in different experiments are always at least in a slightly different states. The modern sequencing machines typically contain a broad range of preprocessing software that already performs many necessary adjustments to the output data. Several types of bias may still remain in the data which need further preprocessing steps. An error source called nucleotide per cycle bias pictures the process where the distributions of sequenced nucleotides changes as a function of the read so that the first nucleotides of the reads are most prominent to errors. Another type of issue is the mappability bias which is caused by varying complexity of different genomic regions. As the complexity of the genomic sequence affects directly the probability of reads mapping to corresponding position, more common sequences have *a priori* a higher probability of finding reads that map to their sequence. In some cases these bias types may seriously affect the following analyses if left unattended. [25]

In addition to these general sources of error, method-specific biases exist also. For instance, in the context of ChIP-seq data a quite recent finding states that ChIP-peaks correlating with high gene expression may in fact contain largerly false signals: the researchers had found that areas of high gene expression exhibit strong ChIP-seq signals even though no protein of interest was actually present [34]. Another bias type related to DNase-seq and especially its application to TF footprinting was recently detected: it appears that the deterministic splicing of the DNA ligase enzyme causes frequent gaps in the DNase-seq signal which can be easily misinterpreted by TF footprint profiling software. This causes serious concerns for only DNase-seq signal based methods of TF binding discovery and might also increase the false positive rates of other TF binding prediction programs making this bias type highly relevant for our analysis [35]. However, no correction method exists for these bias types. Still, it its necessary to acknowledge their existence so that not all results are accepted directly but instead different error sources are carefully considered.

Alignment to a reference genome is a basic step for most NGS data analyses. Several suitable software exist for the task, most commonly known of these being Bowtie and Maq. Bowtie is a fast and efficient aligner that uses the Burrows-Wheeler transform for index construction allowing high performance in sequence querying. Bowtie is capable of aligning both read-space and color space, although the more advanced version Bowtie2 processes data only in read space [36]. Other tools such as the splice junction discovery framework TopHat [37] have been developed on top of bowtie. Maq is another widely used and capable sequence aligner designed originally for Illumina / Solexa read space data but with additions capable of analysis of color space data also [38]. The both aligners allow complex user preferences in, for example, required mapping quality and allowed number of gaps, SNPs, insertions and deletions.

Peak detection

In the scope of this thesis the peak detection is the most crucial part of NGS data analysis. Peak detection is a typical challenge with sequencing types like ChIP-seq and DNase-seq since these datatypes typically exhibit condensed groups of reads in dispersed locations around the genome. To find and distinguish these regions effectively, automated methods are necessary and indeed many such methods have been developed. Here we will quickly discuss two peak detection methods: HOMER and MACS.

MACS or Model-based Analysis of ChIP-Seq data was designed to analyse the output of of short read sequencers. MACS uses a Poisson distribution model to distinguish between signal and background. Using the concept of background is of crucial importance since in every experiment due to errors some reads end up aligning in unexpected positions. Such cases need to be recognized which is usually performed by calculating a background density, often as a long distance average. MACS also takes into account the fact that sequencing of the previously protein bound DNA fragments results in a bi-modal peak. These two overlapping peaks form a characteristic pattern for protein binding and can at best be used to distinguish the binding sequence even in single nucleotide resolution. Generally, MACS has been used widely and is still a very common method in ChIP-seq peak detection. [39]

HOMER (Hypergeometric Optimization of Motif EnRichment) is a motif discovery and next generation sequencing analysis framework first published in 2010. HOMER possesses a great variety of functions and it's peak detection algorithm is capable of detecting ChIP-seq and DNase-seq peaks, histone enrichment regions and detect transcripts from GRO-seq data. The algorithm for ChIP-seq and DNase-seq works by first creating 'tags' at high read count locations. The tags are then

attempted to be shifted center positions of the presumably detected edges of the bimodal peak. By combining density information from different tags the final peaks are formed by combining or masking sequences between nearby tags.[40]

2.5.2 Microarray Data Analysis

Microarray data analysis is a quite well established process due to the crucial role of microarrays in the data-driven development of genomics during the last decade. The fixed parts of every microarray data analysis workflow are preprocessing and normalization, often also followed by differential expression analysis. We now review these concepts in more detail, putting emphasis to those methods relevant to this thesis

Pre-processing

The microarray experiment is naturally prone to all the usual errors of a laboratory experiment. These experiment-related biases like sample treatment and preparation are always difficult in the sense that often not much can be done to correct these early phase errors. Some more minor biases like small differences in sample treatments can usually be handled, however, and they form an important part of pre-processing. In addition, technology related biases are also common: microarray experiment results need to be adjusted according to probe affinities since different probes have different tendency to bind their complementary sequence. In addition, the used fluorescent dyes may induce biases to the data and of course the array itself needs to be carefully constructed and fully functional.

Background correction is a basic preprocessing step where the intensity estimate of the background caused by unwashed fluorescent dyes is subtracted from the probe intensities. Also in the case of two-channel microarrays intensities of the samples measured on one chip are often represented as a log_2-ratio to minimize certain dye-related biases. To increase this further, when measuring biological replicates the samples are often dye-swapped, that is, the dyes used to detect treated sample intensities and control sample intensities are switched in turns. While nowadays also the manufacturers often provide methods like binding affinity profiles for pre-processing of technical biases, despite all efforts many of the mentioned error sources may remain in the data even after pre-processing steps.

Normalization

In addition to pre-processing the data usually also normalization is warranted. In fact, normalization is often considered an inseparable continuum to pre-processing and helps to reduce the effect of still untreated biases even further. Normalization

methods are applied to the data to make it comparable within-chip and between different chips. Many methods of normalization have been developed, especially for the Affymetrix arrays due to their common use. A common normalization procedure for these arrays is the so called Robust Multiarray Average procedure where several consecutive steps skew the distributions of all samples so that the resulting intensity values are proportionate and thus comparable. However, since the microarray data used in this thesis is generated using another technology, we present here a normalization method applicable to our data: quantile normalization.

Quantile normalization is a widely used method where the distributions of expression values from different measurements are made comparable by forcing similar distributions for the values of different samples. The process is illustrated in Figure 2.6. The similar distributions are achieved by sorting the values of each sample and calculating then row averages of the resulting matrix. Here the row average means the average expression of a single gene across all measured samples. The row averages are then returned in places of the original values in order determined by the original value sorting. Since all samples have now the same expression values in various permutations the distributions of the samples are now equal. This is essential before between-sample comparisons can be performed to avoid skewed results. In addition, also within-sample normalization schemes are often applied to the data, for example to make the samples genes comparable with one another. Own methods such as Lowess normalization have been developed for this purpose. However, since in this work we had the advantage to utilize readily pre-processed data we ignore this part of normalization in this discussion.

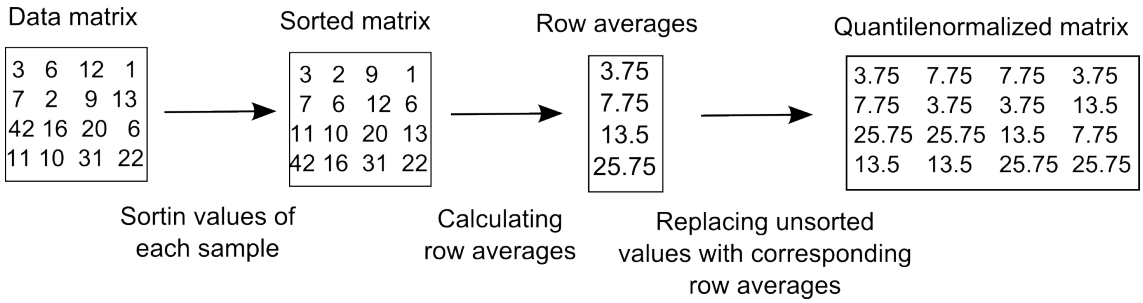

Figure 2.6: The principle of quantile normalization

Differential Expression

Differential expression means the detection of significant difference between expression levels of a target of interest, typically a gene, in two samples. A typical experimental setup compares a sample treated with specific reagent to a control sample with no corresponding treatment. To define the 'significant' difference usually three different conditions must be met. First, the difference in expression must be clear

- usually at least two-fold difference in expression is required. Secondly, the result must be also statistically significant based on some test over treated samples against control samples. A paired t-test is often applied to determine this significance based on differences and similarities in treated sample replicates and control sample replicates. This sets requirements to the whore experimental setting since with a small number of biological replicates a reliable statistical testing cannot be performed and this aspect of differential expression is lost. A third requirement for true differential expression is some underlying biological function which explains the difference. This is of course difficult to determine and is often the final result of the whole experiment.

2.6 Different Ways to Express Transcription Factor Binding

For computational methods transcription factor binding must be representable in an efficiently usable format. Due to the ambiguous nature of TF binding the format must also allow flexibility. In this section a historical method and a state-of-the-art method of expressing TF binding information are reviewed.

2.6.1 Consensus Sequences

Many applications in bioinformatics and biology require the representation of short patterns of nucleotide sequences. Traditionally the binding domains of a TF were expressed as a consensus sequence. The consensus sequence is a stream of nucleotide characters representing a certain domain in the DNA. For instance, a consensus sequence of `TAAGAC` would mean that the TF in question binds only to these specific nucleotides in the DNA. The consensus sequences can be also more ambiguous by replacing some of the bases with characters that signify a choice among several nucleotides. For example the consensus sequence `TANGRS` would be interpreted according to the IUPAC nomenclature for incompletely specified bases as `TA[A,T,C,G]G[G,A][G,C]` where the brackets indicate choice of one nucleotide inside them [41]. This way a set of possible TF binding sites could be easily represented as one consensus sequence. [7]

The usage of consensus sequences suffers from some issues, however. As the TF does not bind only one specific sequence but rather a wider range of sequences the ambiguity of the consensus sequence grows rapidly. For instance, if a TF were to bind often `AACAAT` but also occasionally `TACAAT` and `AACAAA` the consensus sequence would have to be `WACAAW` where `W` represents a choice between `A` and `T`. However, this notation would also allow sequence `TACAAA` which is not among the binding sites of the TF. Thus, the ambiguous representation results into sequences that are unwanted and is therefore a tradeoff between mismatches and precision. Deciding a

consensus sequence is therefore somewhat arbitrary and in general does not represent all the binding domains exactly but only closely enough. [7]

In addition to the tradeoff between precision and errors in a consensus sequence, many TFs typically bind their wide range of different response elements in such a way that each base in the domain has an individual affinity to bind. When in some locations of the domain the base of the response element needs to be always fixed for binding to occur, in other sites the base may vary without preventing the binding but with a contribution to the overall binding affinity. Consensus sequences are, however, unable to take into account this kind of variation in binding affinity when at each location the base is interpreted to be either a match or mismatch, leaving the gray area of partial binding unattended. Due to these limitations of consensus sequences other methods have been developed to represent binding motifs of TFs.

2.6.2 Binding Motifs as Position Weight Matrices

A generally appreciated and more robust solution to representing binding domains of TFs is the position weight matrix (PWM), also known as position specific weight matrix (PSWM). A PWM is derived from a position frequency matrix (PFM) where a frequency of nucleotide occurrences is calculated for each nucleotide at each position of the motif. Thus, the PFM forms a $4 \times N$ matrix where N is the length of the motif and each row corresponds to a different nucleotide. From a PFM the position weight matrix is calculated by diving each frequency count at a specific location by the sum of all frequencies at that location, normalizing the frequencies into weights from interval $[0, 1]$. Mathematically this can be stated as

$$PWM_{i,j} = \frac{p_{ij}}{p_i}, \tag{2.1}$$

where p_i is the sum of all nucleotide frequencies at position i and p_{ij} is the frequency of given nucleotide j at position i. This gives an effortless way to represent a motif of N nucleotides while allowing individually arbitrarily varying nucleotides in the sequence.

A position weight matrix is typically used to represent the binding motif of a TF. Since the binding motifs are often ambiguous in the sense that nucleotides at some positions in the motif may vary between any of two, three or even four nucleotides, it is beneficial to use PWMs to show the motif. Because the continuous nature of PWMs they are superior to the consensus sequences in representing binding: because the ambiguity of bases is not static but instead based on a probability arising from the underlying PFM, much higher resolutions in binding affinities are achieved. Using a position weight matrix each subsequence in the genome can be scored for each TFs binding at single nucleotide resolution. The scoring is performed by select-

ing a value from the PWM corresponding to the nucleotide at that location. The values for the whole sequence are then multiplied together to form one figure that represents the binding affinity between used binding motif and genomic sequence. Due to the definition of PWM values the overall score belongs to interval $[0, 1]$ where 1 represents a perfect match and 0 the worst possible match. The scoring is given as mathematical formulation in Equation 2.2 as

$$PWMS = \prod_{j=1}^{N} \frac{PWM_{i,j}}{b_i}, \tag{2.2}$$

where PWM is the position weight matrix of a TF, N is the length of the matrix, j is the location in the sequence, i is the nucleotide at location j that corresponds to a row in PWM and $PWMS$ is the overall score the sequence receives. Also a normalization term b_i is included here which represents the probability of facing this nucleotide in the sequence. Nucleotide compositions are highly organism specific as the so called GC-content, that is, the amount of G and C bases compared to A and T, varies greatly from one species to another. This has to be taken into account in the scanning by using the presented normalization factor. As can bee seen from the equation the PWMS calculation assumes binding independency of consecutive nucleotides. Attempts to correct the model to take dependence into account have been made but they are not considered in this thesis. [42]

Also another formulation of the PWM score calculation is presented in [42]. This version calculates the sum of log-likelihoods of all locations in the sequence:

$$PWM_{ij} = log_2\left(\frac{p_{ij}}{p_i}\right) \; PWMS_{ij} = \sum_{j=1}^{N} \frac{PWM_{i,j}}{b_i}, \tag{2.3}$$

where variables are as declared in the previous equation. As the sum of log-likelihoods corresponds to a logarithm of product of likelihoods the two definitions are essentially the same with the exception of the latter score is log_2 transformed while the former is not. However, since the former definition scales the results practically to interval $[0, 1]$ and highlights the differences between scores more due to use of multiplication instead of sum, the former representation is chosen to be used in the calculations.

2.7 Previous Approaches to Transcription Factor Binding Prediction

The prediction of transcription factor binding is not a new challenge and attempts to overcome it have been made several times. The different methods can be roughly

divided into two classes. First class of methods considers signatures in the DNase data that indicate TF binding. A binding signature is formed of a area of elevated DNase signal which suddenly drops for a short period suggesting TF binding, and then returns back to its elevated state. As such a signal could indicate binding of any factor, nucleotide sequence of the signal area is afterwards scanned for binding motifs of different TFs. The initial method proposed a greedy algorithm for TF binding signature finding, and was soon followed by other implementations utilizing Bayesian network, hidden Markov model, and Bayesian probability based model which derives its prior probabilities from DNase signatures. [43]

A challenge for these methods is the ambiguity of the TF binding signature: the methods do not distinguish between DNase signals of different TFs. An alternative approach considers this issue by first scanning the genome for putative TF specific binding sites and then introducing the DNase information of the surroundings of these so called candidate binding sites to locate the most probable actual binding sites of the TF. The first such method called Centipede implements a Bayesian mixture model that is adjusted to optimal performance using unsupervised learning and expectation maximization algorithm. Centipede combines data of position weight matrix scanning, sequence conservation, gene proximity information and DNase I hypersensitivity signals to form a comprehensive prediction of putative sites of TF binding to DNA. The developed model was shown to produce admirable results and the authors expected to see it applied to genome-wide prediction studies. [44]

The apparent successor to Centipede published in 2013 is a logistic regression based method called Millipede. The naming convention is due to the reduced number of inputs required by the method even though achieving similar results. The Millipede logistic regression model takes into account the DNase I hypersensitivity information +/- 100bp around the candidate binding site and merges this with predicted binding affinity. The approach extracts the DNase signal and summarizes it to bins in various ways which are then given as input to the method. The Millipede model is formulated as

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{b=1}^{B} \beta_b * D_{b,i} + \beta_{PWM} * PWM_i + \beta_{cons} * CONS_i, \qquad (2.4)$$

where β_0 is the intercept, B is the number of DNase signal bins, $D_{b,i}$ is the sum of DNase signal in bin b at site i, PWM is the score of position weight matrix scanning ath location i (considered in detail later) and $CONS_i$ is the optional sequence conservation score for sequence starting from position i.

The Millipede model was shown to produce as good results as the some orders of magnitude more complex Centipede approach. In addition, with yeast data the Millipede exceeded the performance of the untuned Centipede model showing much

better results making it a method of choice for TF binding prediction at the moment. Thus, Millipede was used in our applied approach as a filtering and comparative tool to evaluate our own findings with a reliable method. [43]

3. MATERIALS AND METHODS

3.1 Datasets

The methods in this thesis use various datasets. All of these have been obtained from freely available sources produced by previously published works. Here a short summary is given of the different data types used in the research and their sources.

3.1.1 DNase I -sequencing Data

The DNase I sequencing data used in the predictions was generated by [45]. The study concentrates to show the connection between DNase I hypersensitive sites and androgen receptor and estrogen receptor binding using hormonal stimulation in LNCaP prostate cancer cell line and MCF-7 breast cancer cell line, respectively. We obtained the DNase I data from sequencing experiment in synthetic androgen (DHT) treated LNCaP cells. According to the published data the cells were treated with DHT for 4 hours prior to the experiment and sequenced with Illumina Genome Analyzer sequencer using manufacturer's instructions in sample preparation[46].

Since the DNase I sequencing data is the most important prediction control element in our binding prediction analysis we used also another data set produced by the ENCODE project. This data was made available by the ENCODE consortium as BED files containing peaks of DNase I active areas. The peaks were calculated by first locating so called DNase hotspots, that is, broad areas of DNase activity and these were further subjected to peak detection algorithm which located local maxima of activity using a fixed length 150 base pair window sliding through the hotspot with 20 base pair steps. These local maxima were then reported in the available BED file as 150 base pair intervals of DNase I activity peaks. The file used in our experiment consists of pooled peaks from all DHT processed LNCaP cell experiments performed withing the ENCODE project. In the case of overlapping peaks pooling was conducted by favoring the peak with higher Z-score calculated based on DNase signal of the non-overlapping parts of the peaks.

3.1.2 ChIP-sequencing Data

For different individual ChIP-seq experiments were used in validations of generated predictions. The ChIP-seq datasets differ slightly from their sample treatments

which enables wider discussion when assessing validation results. The sets are referred to as Sahu, Yu, Massie and Urbanucci. All the sets are generated from LNCaP cells treated with synthetic androgen R1881 before the experiment.

The Sahu ChIP-seq data originates from a recent study AR and GR binding was examined in LNCaP and VCaP cells. The cells in the used sample were treated with 100nM DHT for 4 hours prior to experiment. The resulting sample was sequenced using Illumina Genome Analyzer following manufacturer's protocols. The resulting reads were mapped to genome using Bowtie and scanned for peaks with MACS software. Overlapping peaks from biological replicates were selected with 2% FDR cutoff value. [47]

The Yu ChIP-seq data was originally used in AR subnetwork and TMPRSS2-ERG gene fusion profiling. A dataset with 16 hour R1881 treatment was used. The sample was subjected to 48 hour androgen deprivation before the R1881 treatment The resulting sample was sequenced using Illumina Genome Analyzer standard manufacturer procedure. Reads were aligned using Illumina's ELAND software. Peaks were originally detected using an unpublished Hidden Markov Model-based software so a new peak calling had been performed by the Computational Biology group at Tampere University using MACS peak detection. [48]

The Urbanucci dataset was generated originally to study AR overexpression and its effect on AR-chromatin interaction. One of the measured samples was used in this work. The selected sample was treated with 1nM R1881 for 2 hours and before that subjected to androgen deprivation for 4 days. The used cell line was a special variant of the LNCaP cell line named ARhi, exhibiting native overexpression of AR. Illumina Genome Analyzer was used to sequence the samples, resulting reads were aligned using Bowtie and peak calling was performed with MACS software. [49]

The Massie dataset was generated to support the mapping AR downstream targets. The sample used in this study was treated with 1nM R1881 for 4 hours prior to the measurement. The sample was sequenced using Illumina Genome Analyzer and aligned to reference genome using MAQ aligner. Peak calling was originally performed by MACS and ChIPSeqMini choosing but was repeated by our Computational Biology group using only MACS for comparability. [13].

3.1.3 Gene Expression Data

Gene expression microarray data used in the work was generated with Illumina HumanWG v2 BeadArrays. The two replicate sections of the arrays were treated as technical replicates to minimize small systematic errors noticed to be produced by the method. Raw level bead data was analysed with the manufacturers beadarray software and removing spatial artefacts using an automated method (BASH). The prepared dataset was published in Gene Expression Omnibus (GEO) from where it

was also obtained for this thesis. [50]

3.2 Computational Methods

To make use of the various data sources listed previously several computational methods were applied and constructed to create a framework able to create transcription factor binding predictions and combine these into a genome wide network. This chapter presents the computational methods used in this thesis.

3.2.1 Binding Motif Databases

Several online databases curate and distribute information of transcription factor binding motifs. While some databases charge their usage in this thesis only freely available databases have bee utilized.

JASPAR

JASPAR is an open-access online database of eukaryote transcription factor binding motifs originally published in 2003. JASPAR presents a curated high-quality collection of motifs represented as location based score matrices. In addition to different web based search tools accessible at the database website JASPAR also offers the possibility of database and plain data download. Original JASPAR collection contained 111 scored DNA binding motifs. Later additions have largely increased the size of the database and the fifth, that is, the most current version published very recently in 2014 contains 457 curated motifs [51]. While these have mainly been generated using SELEX experiments also more modern techniques like Chromatin Immunoprecipitation Sequencing have been exploited in determining the motifs. [52]

JASPAR declares to distinguish itself from other similar data sources due to three factors. First the open access is considered a major advantage of the database and it also has a built-in API for easy usage. Most importantly, however, JASPAR's set of binding motifs is non-redundant trying to consist of only the best binding motifs available for each transcription factor. Although this is an issue when considering the modern measurement technologies which data allow the generation of different species specific binding motifs for same TFs, the non-redundancy ensures a clarity and reliability of the presented motifs. [52; 51]

UniPROBE

The universal PBM resource for oligonucleotide binding evaluation or UniPROBE is an open access database of DNA binding domains of transcription factors. Unlike the factors stored in the most widely known databases JASPAR and TRANSFAC

that have been produced with multiple measurement types, the TF binding motif data in UniPROBE has been generated using universal protein binding microarray (PBM) technology. The PBM technology consists of a microarray of short 10-mer nucleotide probes that are designed to resemble all possible permutations of nucleotides. The probes are double stranded and TFs of specific type are allowed to bind the sequences. These bindings can be detected and thus the bound sequences can be deduced. This measurement gives a high-throughput way to map all possible binding motifs of short binding motif TFs. [53]

The original UniPROBE database contained PBM derived non-redundant binding motifs from 175 TFs from various species. The database has been updated with several of publications and at the moment contains 11 different publications. The collection currently consists of 406 TF binding motifs from several species including human. [54; 55]

TRANSFAC

The TRANSFAC database initially published in 2010 provides a commercial option to the open-access TF databases [56]. TRANSFAC is a resource of eukaryote TFs, their binding sites and binding motifs represented as position weight matrices. The database is continuously curated and its features are frequently updated. A restricted and outdated free version of TRANSFAC from year 2005 is also provided [57]. The public database is, however, severely inferior to the commercial version: the free version contains 398 PWMs while the commercial one claims to contain even 5551 PWMs. The commercial database also promises advanced data search and analysis tools and a possibility to data download [58].

3.2.2 PWM Scanning

Position weight matrix scanning was performed using a modified version of PSWM scanning software by a Tampere University Computational Biology group collaborator Harri Lähdesmäki. The original code written in MATLAB implements a simple DNA sequence scanning with a given $4 \times N$ PWM using the product based calculation method presented in equation 2.2. The algorithm also calculates a background score based on nucleotide frequency in the region to-be-scanned, thus normalizing the resulting scores in respect of the bias caused by uneven probability of different nucleotide occurrence. In addition, to extract significant scores the algorithm performs a background scan using a 100 fold longer randomized sequence than the actual sequence-to-be-scanned. The scores from the inspected sequence are then thresholded so that all DNA locations with a score equal or above a fixed threshold highest scores from the background scan were accepted as so called candidate

binding sites.

The performance of the scanning was first validated by scanning the 1000bp upstream sequence of gene ELK4. Confirmed binding sites of AR had previously been reported in the promoter proximal region of ELK4 at positions $-167/-153$ and at $-481/-467$ counting from the transcription start site of the gene [59]. The peaks of the scanning were plotted to Figure 3.1. As can be seen the implemented PWM scanning finds relatively high scores very accurately at these locations. Even though the left peak seems quite modest in size it was separately confirmed that all the visible peaks in the plot are at least one order of magnitude larger than the remaining peaks. This finding shows arguably that PWM scanning can be used as an effective prediction tool for finding TF binding sites. It is worth noticing, however, that in addition to the two confirmed peaks the prediction finds a few other peaks as well. These can well be unannotated true binding sites of the AR but also just random bits of sequence happening to have a permutation that resembles the AR binding motif. Indeed, the distinction between true binding sites and false positives is one of the hardest tasks in PWM based TF binding prediction.

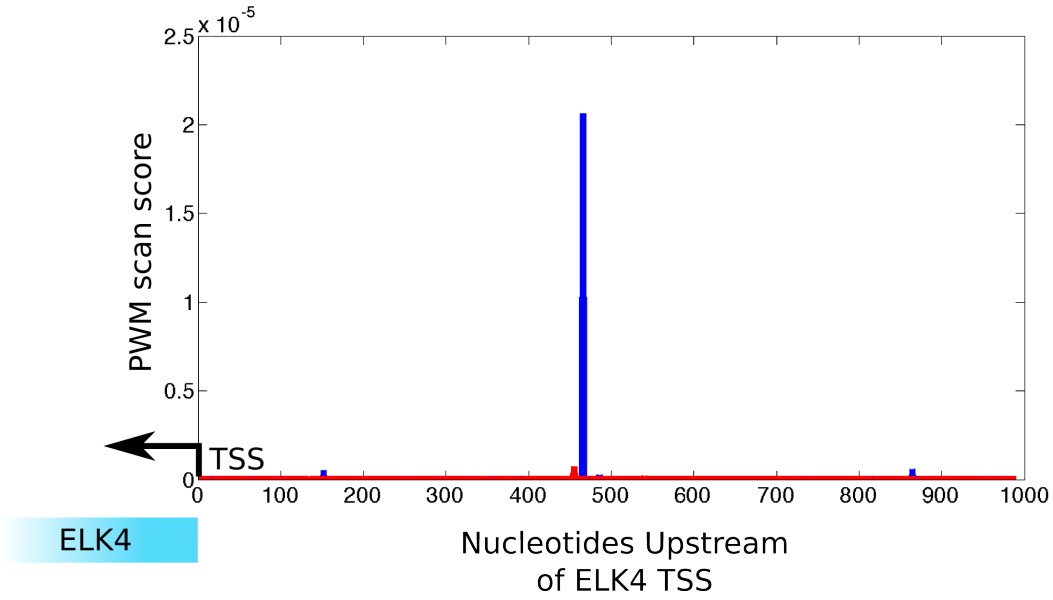

Figure 3.1: PWM scan of ELK4 1000kb upstream sequence

The suggested threshold of 0.1% was selected also to our analysis after examining AR binding upstream of gene TMPRSS2 transcription start site. The Figure 3.2 shows a log_2 transformed histogram of the PWM scan scores bor randomized background scan (dark blue) and scan of a 100kb upstream sequence from TMPRSS2 gene (cyan). The light blue effect shows the areas of the histogram where the two plots overlap. As the PWM scan results were in range $[0, 1]$, after the log-transform the scale becomes $(-\infty, 0]$ where larger value is better scan result. As can be seen

from the figure, even though the TMPRSS2 upstream sequence scan produces 100-fold less scores than the broader background sequence scan, the right tail of the sequence scan seems to be thicker than the tail of the background scan. This gives more confidence in the AR PWM motif we are using since it seems that scanning actual biological sequence with the motif produces more high score hits than scanning completely random sequence. This corresponds to the biology of the scanned sequence: the TMPRSS2 upstream sequence is expected to contain several binding sites for AR and at least according to these histograms it seems we are able to locate some of these sites.

The significance threshold filtering was modified to express the output significance as pvalue based on location in the background threshold, thus calculating for each gene i at location j

$$pval_{i,j} = \frac{\#(PWMS_{i,j} > bg)}{\#bg}, \tag{3.1}$$

where $PSWM_i$ is the PWM scan score for the ith gene at some location j, bg is the list of background scan scores and $\#$ signifies the number of elements. Now the previously formed threshold of rejecting 0.999% least significant genes turns into pvalue threshold $pval = 0.001$.

The filtering threshold for significant hit in this example becomes in log_2 scale -24.1. This seems to be a rather strict criterion judged from the figure but it was selected since it certainly rejects the bulk of low scores and should severely restrict the number of false positives. Even though different PWMs have varying lengths the background scanning is assumed to ensure that the threshold calculation is scalable to other transcription factors also and thus acts as a reliable criterion to separate significant scores from non-significant ones.

3.2.3 DNase Peak Detection

Peak detection of the Housheng et al. dataset had been performed by the Tampere University Computational Biology group researchers previously. The method used was HOMER peak detection. The peak detection for ChIP/DNase-seq peaks had been used for our DNase I sequencing dataset with peak size estimation 185bp. From the detection results it can be seen that the estimation is only loosely followed since the resulting observed peak size on average is 355bp.

3.2.4 Prediction Efficiency Validation

Predictions generated with different methods were validated by examining the predictions of AR binding using the ChIP-seq dataset as ground truth. For simplicity, all comparisons were performed in chromosome 20 and assumed robust enough to

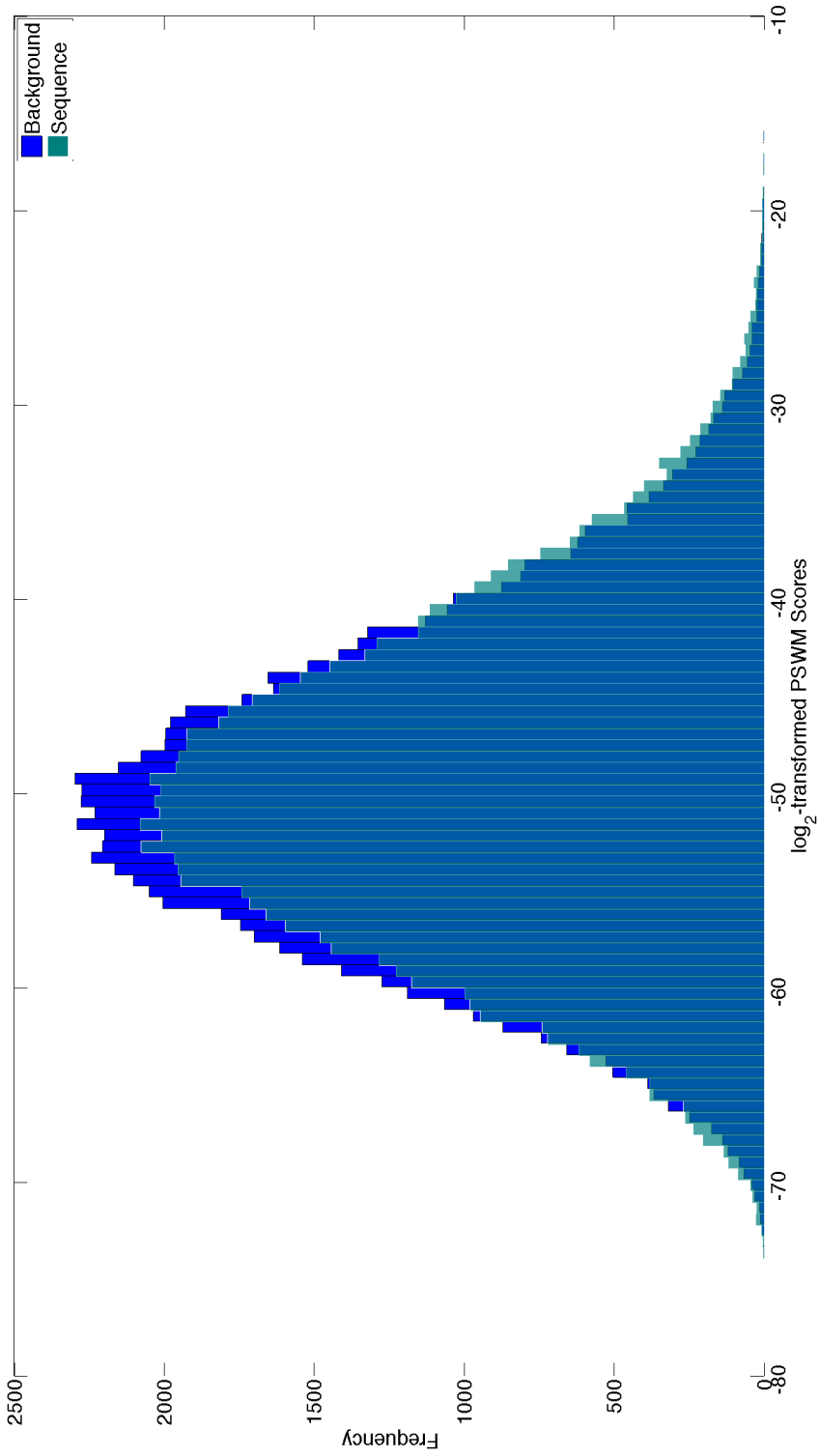

Figure 3.2: Histogram of TMPRSS2 upstream sequence scan againts scan of randomized background

be applicable to whole genome level if necessary. ROC curves were used to represent effect of thresholding PWM scores. Here we hypothesized that filtering the low PWM scores would reduce the number of false positives as gradually only the sites with strongest affinity remain. For the ROC curves the validation ChIP-seq peak set was prepared in similar manner as described in the Millipede publication: Positive TF binding set was defined to be the overlapping hits of ChIP-peaks and predictions while the negative TF binding set was defined to be the set of candidate binding sites not falling into ChIP-peaks. In the context of ROC curves this construction forgets the existence of possible ChIP-seq peaks that are not detected by predictions and shows only the changing performance in finding the detectable peaks when the filtering criterion of PWM score is changed. To consider also the fraction of missed ChIP-seq peaks this is reported separately with every approach along with discussing the number of true positives and false positives in the case of unfiltered PWM scores.

3.2.5 Gene Expression Enriched Network

To form most accurate picture of which regulation connections are present in the cells, gene expression data from LNCaP cells was applied to the constructed network. This was done in two separate phases: first, static state gene expression data with no androgen stimulation was used to filter out all regulation connections between very low expression genes from the network. This forms a greatly reduced new network which pictures more accurately the regulation present in the cell in steady state. In addition, as a second phase, DHT stimulated gene expression data was used to uncover the downstream regulation patterns of AR in high detail. The differentially expressed genes at time point 4 hours were calculated from the time course data and these were used to combine the findings of predictions and actual measured differential expression, and as a proof a small AR downstream network was profiled.

3.2.6 Predicted Network Inference

The network of regulation connections was constructed from the predictions by considering the predictions 250kb upstream of gene TSS possible regulators of the gene. Locations of these predictions along with the TF and target gene were saved to a file in BED format. For visualization this format was still flattened in boolean yes/no fashion to a list of connections between all TFs and genes. This list was then used to visualize the predicted networks using program Cytoscape, a commonly used network visualization tool.

4. RESULTS AND DISCUSSION

4.1 Assessing the ChIP-seq Validation Reliability

To be able to validate the results of different prediction approaches the quality of the ChIP-seq data used is of vital importance. To ensure the robustness of performed validations four different ChIP-seq datasets were used independently to validate all PWM scanning results. A summary of ChIP-seq peaks inferred from these datasets is presented in Table 4.1. As can be seen the sets contain quite different amounts of detected peaks. The sets had, however, received slightly differing treatments prior to the experiments which without doubt explains greatly this variance. Another explanation is the used peak calling. Even though the peaks of all datasets were detected with same software it is probable that different parameter settings resulting in different numbers of peaks.

To validate the datasets in the sense of PWM scanning the sequences corresponding to reported peaks in chromosome 20 were extracted from the UCSC human genome version hg19. These sequences were subjected to PWM scan using the binding motif of AR and maximum scores were calculated from each segment. Assuming the filtering criterion calculated when examining AR peaks in upstream sequence of TMRPSS2 stays relatively same, the number of maxima larger than the threshold $pval = 0.001$ corresponding to PWM scan score $5.7403e - 08$ were calculated. These sums were collected to Table 4.1.

If a ChIP-seq peak covers such a genomic sequence where all PWM scores are below the set threshold, this peak is impossible to detect for any PWM scan based method. As can be seen from the table the numbers of detectable peaks against all reported peaks differs quite largely in the case of stricter filtering with $pval = 0.001$. The Massie and Urbanucci sets seem to match quite well areas of detectable peak scores. However, the large number of peaks in the Yu set on areas where

Table 4.1: A summary of the ChIP-seq datasets used in validations

	Sahu	Urbanucci	Massie	Yu
Total number of peaks	16971	2399	11785	35363
Peaks in chr20	449	57	233	758
Detectable peaks, pval=0.001	278	39	181	489

only relatively low PWM scan scores were observed raises doubts concerning the experiments success: approximately 250 peaks were located on areas of low PWM scores which is a significant percentage of all the Yu peaks. While difficult to prove, this the possibility of bad quality dataset naturally exists. There can easily be issues in the ChIP-seq sample preparation, measurement, read processing or peak detection along the way and it is practically impossible to tell whether or not something unwanted has happened to the data since it is generated and analysed by us. Thus, using another ChIP-seq dataset for validation, or perhaps a combination of sets would result in a valid.

Another likely explanation is that the unobserved peaks are caused by indirect binding through protein-protein interactions. The same holds true for the other datasets as well This indirect binding could well be observed in the ChIP-seq experiment when the whole complex is cross-linked to DNA and gets selected by antibodies due to AR being one of its subunits. Indirect binding is, however, difficult to prove and since the used methods are require DNA binding for detection the indirect binding and its effects are unfortunately outside reach of this study.

4.2 PWM Scanning

PWM scanning was performed using each of the 672 TF motifs extracted from the publicly available JASPAR and UNIPROBE. The sequences 250kb upstream of every gene annotated in UCSC genome version hg19 were first extracted from the hg19 primary assembly to a FASTA file based on annotated gene coordinates. Even though the initially scheme when planning the experiments was to scan all of this sequences with all motifs, this proved quickly too time demanding for the scope of this work. Especially the 100-fold background scan phase increased the predicted runtime of the PWM scan to several months resulting into rejecting this original plan. Instead, the information from DNase sequencing was used first to filter out areas of closed chromatin in order to reduce the number of sequence-to-be-scanned.

4.2.1 Naive DNase Filtering

The first attempt to utilize DNase signals calculated from the DNase I sequencing experiment was performed using the simplest possible approach. The sequenced reads were transformed into a bedgraph file with reverse and forward strand reads counts summed together to form a DNase-seq signal. The approach then applied a fixed threshold to the signals stating that areas with signal higher than the threshold are the areas of interest with accessible chromatin. A histogram of the DNase-seq signals in chromosome 20 was used as a rough tool in determining a fitting threshold. Judged from the histogram it became clear the bulk of the DNase-seq signal values

Table 4.2: Naive DNase-seq signal based PWM scan

	Sahu	Urbanucci	Massie	Yu
Overlaps	38	8	18	31
Percentage of detectable	13.7	20.5	9.9	6.3
True Positives	84	25	41	66
False Positives	190	249	233	208

is near zero or zero signifying the majority of the DNA is inaccessible to TFs. After this examination the threshold was set so that signals with values larger than 95% of all signals were judged significant. Calculated as a signal threshold and rounded to integer, as the signals were integers also, this gave a threshold value of 4. As seen in the histogram the threshold seems to quite effectively filter out the bulk of low signals while the interesting high signals remain.

This scan yielded in total $3,448,766$ binding sites for all TFs which corresponds to approximately $5,100$ binding sites per TF on average genome-wide. When considering the scanned sequence length was still roughly 41,000,000bp for each of the PWMss this number seemed surprisingly small. Indeed, when examining the binding predictions of AR in chromosome 20 only 277 predicted binding sites were found in the whole chromosome.

Results of the scan validations with all ChIP-seq datasets is presented in Table 4.2 In the table the first row shows the number of explained ChIP-seq peaks, the second tells the percentage of explained peaks among all detectable peaks, the third shows number of predictions overlapping ChIP-peaks and the fourth number of predictions missing all ChIP-peaks. These validations were implemented by calculating overlaps between predictions and ChIP-seq peaks allowing partial overlaps but no tolerance. As can be seen from the table the overlaps between predictions and validation peaks are scarce. It seems that every one of the validation sets are explained rather poorly by the predictions: only few predictions overlap with ChIP-peaks while still majority of the predicted binding sites are false positives.

Also a ROC curves were plotted to determine whether filtering of the predictions according to PWM score results in fewer false positives and better precision. The ROCs are shown in Figure 4.1. As can be seen the thresholding with PWM scores does not seem to result in better performance in this case. Instead the few true positives predicted seem to get filtered out even faster than the false positives, suggesting that at least after some threshold the PWM score might not be in crucial role in determining the true binding sites.

Two different explanations were devised for this behavior. Either the DNase-seq signals used do not overlap the ChIP-seq validation peaks in many positions at all, or the method the DNase information is used is not optimal. The overlaps between

Table 4.3: Overlaps between DNase-seq signals and ChIP-seq peaks

	Sahu	Urbanucci	Massie	Yu
Overlapped peaks	388	52	197	592
Percentage covered with DNase	86.1	91.2	84.6	78.1

the Housheng Dnase-seq dataset and ChIP-seq peaks was inspected and it was noted that while not all peaks were covered by DNase signal the majority of them showed at least partial DNase-seq signal overlaps. The overlap percentages are presented in the Table 4.3. The overlap percentages are quite high indicating that almost all peaks are throughly covered with DNase signal. This suggests that majority of the peaks be detectable also according to DNase-seq signals. Thus, the issue with prediction precision seems to lie elsewhere that in insufficient DNase-seq coverage.

The second apparent error source is thus the used method itself. Since very low numbers of peaks were detected by predictions - for example with the Sahu data only 8.5% of the ChIP-seq peaks - while generously over half of the peaks should be detectable, it seems that the detection method indeed is too inaccurate. This was looked in more detail by examining the DNase-seq data using the Integrative

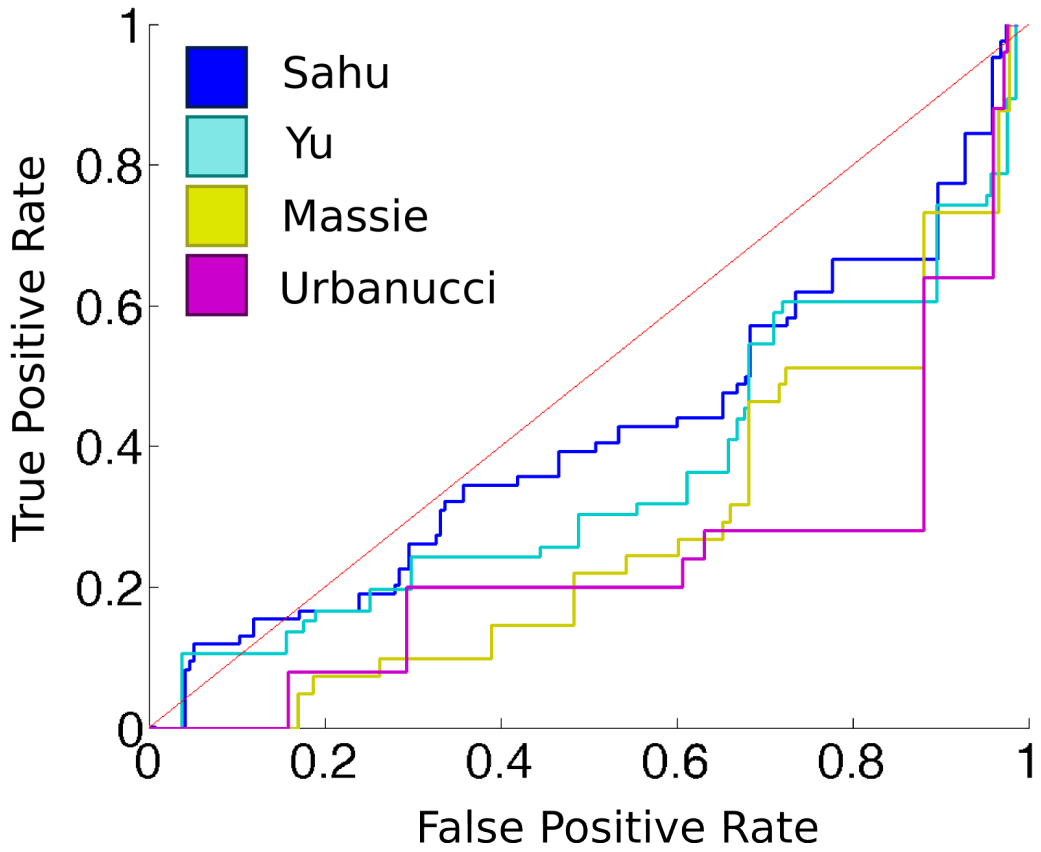

Figure 4.1: ROC curve of naive DNase-seq based PWM scan

Genomics Viewer (IGV) software and showing the predicted binding sites and the DNase-seq signals resulting from the thresholding. We noticed that the DNase-seq sequencing depth and the nature of the data both contribute to the fact that the thresholded DNase-seq signal areas observed were short and fragmented. The limited sequencing depth and on the other hand the fact that there is a gap in the DNase-seq signal at protein binding sites resulted in frequent fragmentation of the DNase-seq signals since the fixed threshold allowed no flexibility and an area considered open chromatin was cut every time the signal descended below 5. The short fragmented scan sequences allowed unrealistically short areas of DNA to be scanned and large amounts of candidate binding sites were lost. In addition, the thresholding process also eliminated some of the most promising binding sites since in many cases at sites clearly resembling protein-bound DNase-seq signal when looked by plain eye via IGV, had so low signal gap that the sequence was split in two just at this critical location. Thus, this first attempt was named naive DNase filtering due to its simple fixed threshold based usage of DNase-seq information and as a following step a more refined approach to utilize the DNase-seq information was devised.

4.2.2 DNase Peak Scan

The refined approach to improve the quality and length of PWM scanned sequences was to use peak detection with the DNase-seq data to obtain the areas of relatively strong DNase-seq signal as broad peaks. The detected peaks were used to extract corresponding DNA sequences which were subjected to PWM scanning using all TF motifs as before. The high score sites were saved based on similar 100-fold randomized background scan and $pval = 0.001$ criterion. The scanned sequence was extracted from the UCSC genome version hg19 directly and contained no gene proximity information. After the scanning and thresholding the actual network was then inferred from the predictions based on their location using hg19 gene annotation coordinates.

The validation results are collected in Table 4.4. We notice that the refined approach based on DNase-seq peaks produces much better results in terms of detected ChIP-peaks than the previous naive method. Especially the percentage of detected peaks of Sahu and Urbanucci datasets are promisingly high. On the negative side, the amounts of false positives are quite high indicating a large number of binding sites being predicted in locations not validated to be AR binding sites in the cells' measured state.

Also ROC curves shown in Figure 4.2 were plotted to picture the behaviour of the predictions when PWM score is thresholded. The Figure shows that with higher PWM score binding sites the ChIP-seq peaks are detected with slightly less false positives compared to the basic threshold. Still, the enhancement is quite modest

Table 4.4: DNase peak based PWM scan

	Sahu	Urbanucci	Massie	Yu
Detected	209	31	98	228
Percentage of detectable	75.2	79.5	54.1	46.6
True Positives	470	58	237	492
False Positives	5035	5447	5268	5013

and like in the naive case, it seems that no significant benefit in detection quality can be drawn from the raising of PWM score threshold.

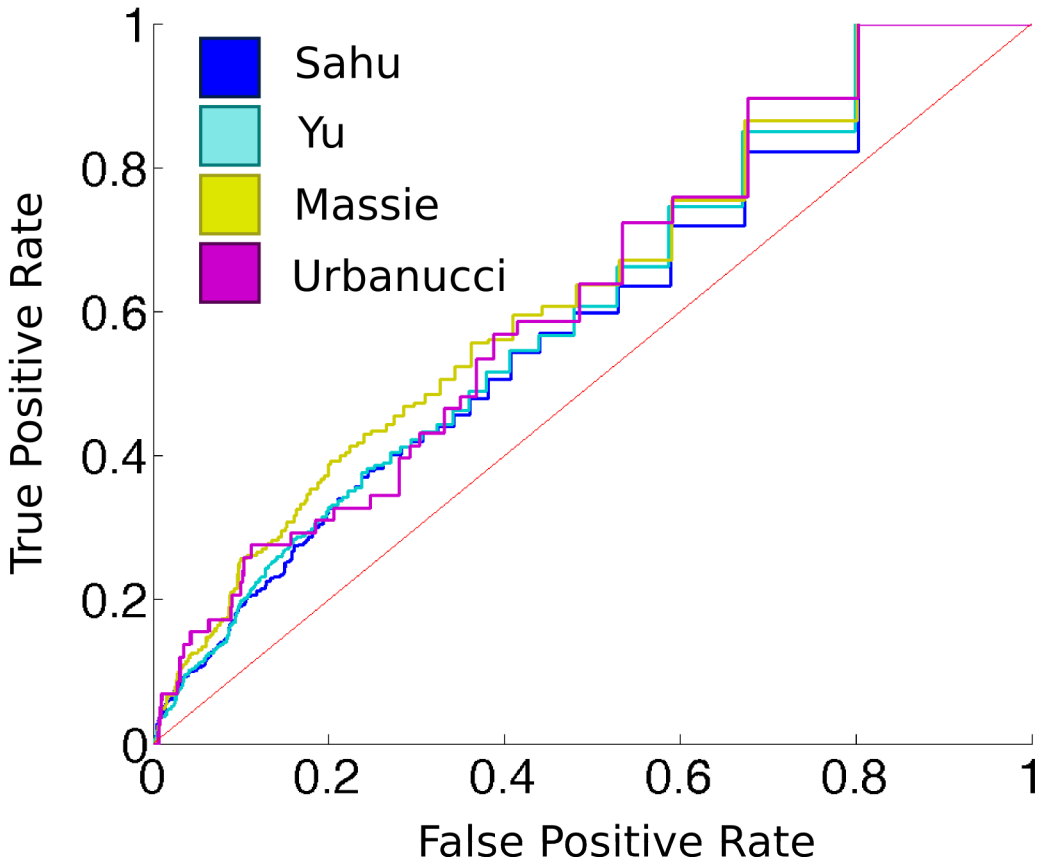

Figure 4.2: ROC curve of DNase-seq peak detection based PWM scan

4.2.3 ENCODE DNase Data

The ENCODE DNase-seq experiment was used to extract DNA sequence corresponding to the area of observed peaks. Results of this approach validations are collected to Table 4.5. The results seem to be very similar to those using Housheng DNase data. This suggests the quality and composition of the two datasets are

Table 4.5: ENCODE DNase peak based PWM scan

	Sahu	Urbanucci	Massie	Yu
Detected	213	32	101	234
Percentage of detectable	76.6	82.1	55.8	47.9
True Positives	487	60	246	511
False Positives	5264	5691	5505	5240

similar and truly very minor differences in prediction statistics can be observed. On average the ENCODE approach seems to predict a few more ChIP-seq peaks. On the other hand, however, the ENCODE DNase-seq set contains more peaks which can be seen mostly in increased false positive predictions compared to corresponding Housheng predictions.

ROC curves plotted by thresholding the PWM scan score are shown in Figure 4.3. Interestingly, the figure is almost identical to the previous case with the Housheng data. This probably tells from the similarity of the two DNase-seq signal sets. Since we observe predictions with almost similarly behaving PWM scores it seems the likely that the predicted sites are in fact largely the same even though the DNase-seq data is different. The two DNase-seq datasets should even be expected to be quite similar since after all they are measured from the same cell type that received similar treatments. As the numbers of true positives and false positives correspond to each other rather well between Housheng and ENCODE datasets, these findings together suggest that in these different experiments the underlying DNase hypersensitivity patterns are similar.

4.2.4 Millipede Filtered Peak Scan

Although achieving fair detection sensitivity in the previous approach both with Housheng DNase-seq data and ENCODE data the large fraction of false positives remained an issue. To answer this problem an attempt to utilize Millipede TF binding prediction framework was performed. Locations of the candidate binding sites found in the previous phase were used to collect corresponding DNase-seq signals in single nucleotide accuracy. The PWM scan scores of these candidate binding sites along with the DNase-seq signals were subjected to unsupervised Millipede prediction using the signal mode "M2" which was suggested in the original publication. The resulting Millipede scores were used to filter the candidate binding sites, choosing only those sites with the best Millipede score.

Since the Housheng and ENCODE datasets were noticed to perform almost identically, only one of these was chosen for the Millipede filtering. Having produced

Table 4.6: DNase peak based PWM scan

	Sahu	Urbanucci	Massie	Yu
Detected	101	15	59	115
Percentage of detectable	36.3	38.5	32.6	23.5
True Positives	151	19	89	156
False Positives	985	1117	1047	980

slightly less false positives the Housheng dataset was selected. Millipede filtering was observed to lower significantly the number of predicted binding sites in chromosome 20 for AR. For example, with the Sahu dataset the number of predicted sites was reduced from 5506 to 1137. The validation results of Millipede filtered binding predictions are collected to Table 4.6.

We notice that the Millipede filtering indeed functions as a tool in reducing the false positives as the number of these drops roughly to one fifth compared to the unfiltered predictions. However, also the number of true positives drops as the

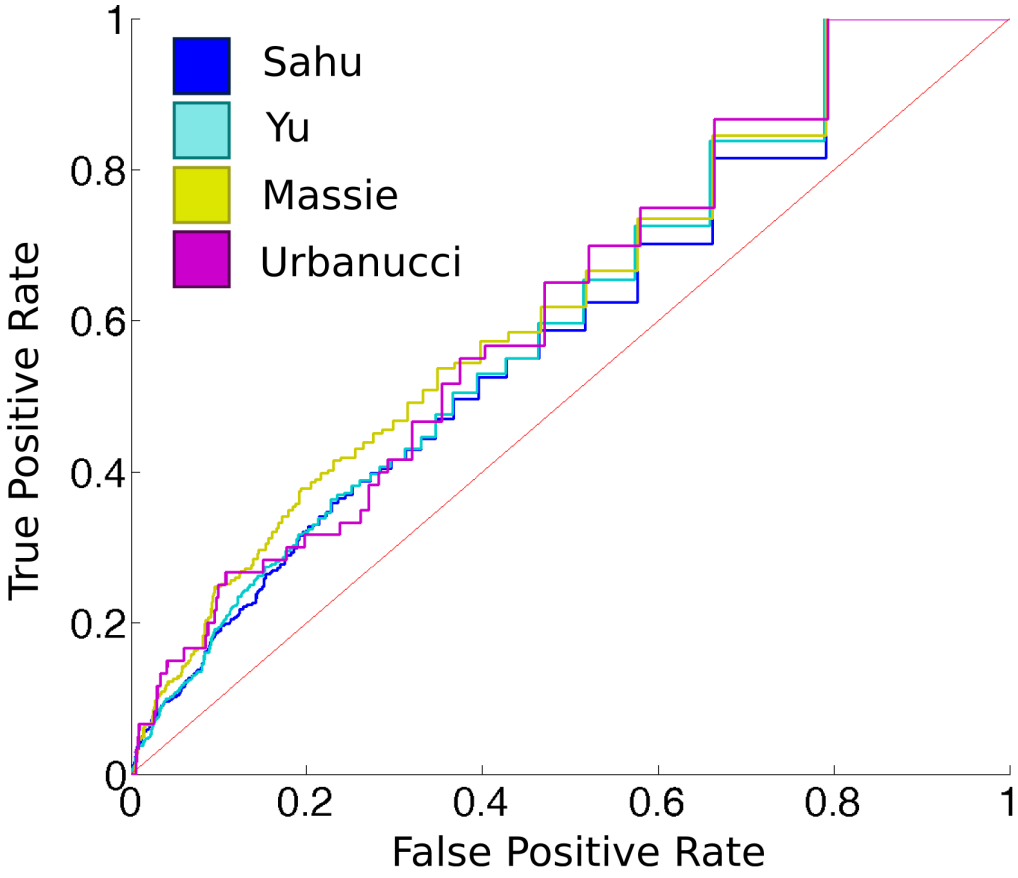

Figure 4.3: ROC curve of ENCODE DNase-seq peaks based PWM scan

method filters out also correctly predicted binding sites. The ratio of the reduced true and false positives seems to give a slight positive signal for the filtering: for example with the Sahu dataset, while the number of correctly predicted ChIP-seq peaks reduced approximately 56%, simultaneously the number of false positives was reduced 80%. This indicates the Millipede logistic regression model seems to be able to reduce the number of false positives more than it loses real binding sites. However, the benefit achieved by this method is still modest in the sense that we do lose almost half of the true positives in the process raising questions if the Millipede-assisted filtering worth its while.

Also a set of ROC curves presented in Figure 4.4was again drawn by thresholding the PWM scores. The ROC curves resemble largely the corresponding curves before Millipede filtering. The Urbanucci dataset behaves slightly differently in validation than before yielding higher area under ROC curve. Also the form of the curves is lightly affected by the filtering: at FPR 0.5 other than the Urbanucci set show some worse performance than before filtering. Even though the changes are minor they prove that the Millipede filtering seems to indeed pick binding predictions based on the DNase signature, resulting into a differently distributed set of filtered PWM scores.

4.2.5 Tolerance Inspections

To see whether the missed ChIP-seq peaks were only close misses or far away from our predictions the candidate binding sites were next allowed different tolerances. The tolerance was added to each of the predicted sites on both sides so the binding site was widened two times the tolerance length. The resulting amounts of explained ChIP-seq peaks were collected to Table 4.7.

Table 4.7: AR prediction covered ChIP-peaks in chromosome 20 with varying tolerance

	+/- 0	+/- 50bp	+/-100bp	+/-500bp
Housheng	227	227	237	277
ENCODE	213	231	240	280
Housheng Millipede	101	111	115	151
ENCODE Millipede	90	99	103	132

As can be seen from the table, expectedly the number of detected ChIP-seq peaks steadily increases as tolerance is let grow. In general, all experiments seem to benefit from the increased tolerance but the increase in detected peaks is mostly quite small. Increasing the tolerance slightly by +/- 50bp the most advantage is gained by the ENCODE approach without Millipede where the number of detected peaks rises almost 20 pieces, while the Housheng data without Millipede filtering

does not detect any new peaks. Notably, the Millipede unfiltered approaches reach very high prediction fractions as the tolerance is risen to +/- 500bp. However, this number is unrealistically high, which is given away already by the fact that the ENCODE approach detects 2 peaks more than were judged detectable based on the PWM scanning score threshold previously. That is, the tolerance turns the intervals so wide that they start to overlap some of the undetectable peaks also.

Generally, it seems that some of the detectable but missed ChIP-seq peaks are located in the proximity of predicted peaks and could be near misses. The fraction of these cases is still so small, though, that with a realistic tolerance the increase in prediction performance is only minor. Also from Table 4.8 can be seen that no significant descent is observable in the number of false positives when increasing tolerance to +/- 100bp. This suggests the predicted binding sites deemed as false positives are not in the proximity of the true positive sites but instead divided more evenly across the whole chromosome.

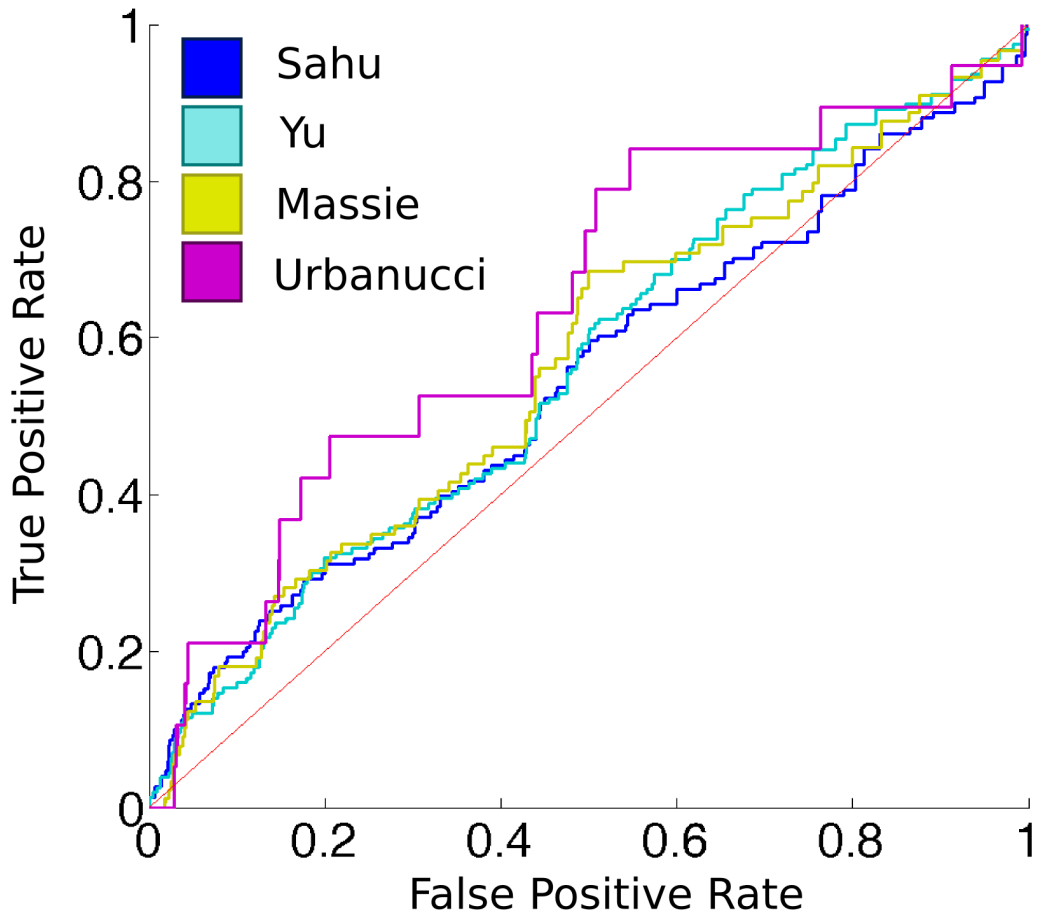

Figure 4.4: ROC curve of Millipede filtered Housheng DNase-seq peaks

Table 4.8: Predicted AR binding sites at different tolerances

	TP, 0bp	FP, 0bp	TP, 100bp	FP, 100bp
Housheng	535	4870	597	4808
ENCODE	448	5264	619	5132
Housheng Millipede	151	985	182	954
ENCODE Millipede	133	871	157	847

4.2.6 Adjusted Thresholding

To determine truth about the suspicions that too strict PWM score threshold had initially been used the PWM scanning was repeated for the AR binding motif using a filtering threshold of $p = 0.01$ visualized also in the previously shown histogram in Figure 3.2. The sequences-to-be-scanned were selected from the detected DNase-seq peaks in Housheng data. The results show that on genome-wide level a staggering amount of $2,072,846$ candidate binding sites were approved, $47,755$ of these in chromosome 20.

As based on the vast number of predictions it seemed unprobable that this approach would provide major benefit, only one ChIP-seq dataset (Sahu) was used for a quick validation experiment. With the lowered filtering criterion in mind the DNA sequence of the ChIP-seq peaks was again inspected for PWM scores and it turned out that with the criterion $p = 0.01$ which corresponded to PWM score $6.3 * 10^{-}10$ a total of 445 ChIP-seq peaks contained at least one binding site with PWM score this high.This means that only 4 of the ChIP-seq peaks remain undetectable and the new prediction set should be able to find almost all of the ChIP-seq peaks, at least if the DNase-seq signals are suitable.

The almost fifty thousand predictions were matched to ChIP-seq peaks which resulted in 310 detected peaks and in total 3425 true positive predictions while producing 44329 false positive predictions. These results when compared to the earlier results give an unpleasant picture of the effect of lowering the PWM scan threshold: as the number of true positives grows about 7 times larger compared to the results of the stricter filtering, the number of false positives is approximately 9 times greater than in the corresponding strict filtering results. This seems to suggest that the lowered threshold adds more false positives to the predictions than what can be gained in true positives questioning the usability of the lowered threshold. A ROC inspection was ignored since it has been clearly demonstrated that the PWM score thresholding does not have a positive effect on prediction accuracy.

To see if advanced filtering would show any benefits with the increased prediction size, Millipede filtering with same settings as before was run. The approach reduced the set of candidate binding sites to $12,292$. The number is approximately 10

times the corresponding number of predictions with the stricter PWM scan filtering criterion. This means the Millipede filtering as also the whole DNase-seq peak detection based PWM scan seems to be linearly scalable as no extra benefit is gained from the filtering even though the number of candidate binding sites rises. The resulting set of candidate sites predicted 268 of the 445 detectable ChIP-seq peaks meaning a better result than the analyses with the stricter criterion. Also the fraction of detected peaks found after and before Millipede filtering is now higher than in the stricter criterion (where we had 101 and 227 detected peaks, respectively) suggesting that the Millipede filtering might be able to extract some more correct information as the number of predictions increases. The cost of this in terms of false positives was 11308 while 983 predictions were categorized true positives.

In general, it must be stated that no notable benefits were to be gained by using a looser PWM score filtering criterion. The number of predictions rises this way to a highly unrealistic amount and as no efficient way of filtering these for the correct positions, the results are hardly more usable than those of the stricter filtering criterion. Still, it has to be admitted that this is the only way of being able to detect almost all of the ChIP-seq peaks. However, questions whether we even should detect the peaks in these quite low-scoring areas remains open.

4.2.7 Comparison with Native Unsupervised Millipede

Altogether the so far observed results clearly indicate it is not easy to distinguish the "true" binding sites of the AR indicated by the ChIP-seq peaks among false positive predictions. Indeed, it was considered whether this is even possible at all. The sites deemed false positive are according to the method sites of quite high affinity to AR and on areas of open chromatin. As discussed in the Chapter 2, the binding of the TFs is far from a deterministic process but instead the binding frequency depends on numerous factors. Interestingly it was also discussed that the binding of TFs is not even directly related to the concentration of the TF in the cell. Thus, it could be fairly possible that due to some more complex protein-protein interactions some of the open high affinity binding sites are just left empty: they could easily have another TF or other protein nearby which inhibits binding of extra elements to their general proximity. To see whether this is truly the case the whole sequence of chromosome 20 was subjected to basic PWM scanning and all the resulting PWM scores were passed along to Millipede along with corresponding DNase-seq signals. Since Millipede had been shown to perform admirably in previous studies even in its unsupervised mode the predictions generated by Millipede could be considered 'reliable', and it's results would provide more information of the behavior of these different methods.

The resulting scores from the Millipede run were thresholded allowing only the

Table 4.9: DNase peak based PWM scan

	Sahu	Urbanucci	Massie	Yu
Detected	33	1	3	16
True Positives	441	13	45	205
False Positives	2017	2445	2413	2253

highest scores to be interpreted as candidate binding sites. The choice of this threshold was somewhat arbitrary, but the resulting number of 2458 predictions using the strictest score did not encourage to selection of a lower criterion. These predictions were compared with the ChIP-seq peaks and the results were collected to Table 4.9. Surprisingly, with this strict filtering the numbers of correctly predicted ChIP-peaks are very low for all datasets. When examined more closely the high scoring Millipede predictions were found to consist of the highest scoring PWM scan results which had also adequate DNase-seq signal, that is, the high PWM score seemed to be emphasized in the results and the DNase-seq signal was used more loosely. This is the case using the suggested default parameters with the Millipede logistic regression model. The weights could likely be adjusted so that at least some differences in the predictions could be observed but this tuning was left outside of the scope of this thesis.

The Millipede method, using suggested binning and default weights for DNase-seq signals and PWM scores produces a high number of predictions that hit relatively small areas in the genome, as for example in the case of the Sahu dataset, 441 predicted binding sites were condensed in only 33 ChIP-seq peaks. We hypothesize this is due to different approaches in the DNase-data usage: when confronting an island of high DNase-seq signals the Millipede is free to select each site with high PWM score as a very probable binding site, thus giving it a maximum score. However, our DNase-seq peak based method is forced to PWM scan only the sequences of relatively short peaks which apparently yields smaller number of high scoring PWM scan results. This leads to smaller number of predictions with wider location range. It seems thus probable that the Millipede prediction could lead into most accurate coverage of the ChIP-seq peaks if the Millipede result score threshold was lowered enough. This would be done, however, with a price of false positive explosion.

4.3 Gene Proximity Filtering

The binding sites of TFs occur frequently in gene promoters and promoter proximal regions since these are the main areas of transcriptional control. With this in mind the number of false positives predictions was inspected by thresholding the predictions based on gene proximity. Distances of each predicted binding site were

calculated to each genes TSS and the minimum distance was selected as a 'gene proximity' score.These scores were then thresholded to produce the ROC curves showed in Figure 4.5. When compared to PWM scan score thresholding the proximity filtering produces clearly different results. It seems that the gene proximal regions are covered with nearly equal amounts of false positives and true positives providing no help in their distinction. As the distance to gene TSSs grows it seems to contain some information capable of predicting more true positives than unthresholded version. In addition, interestingly on long distances to TSSs the predictions seem to be exclusively false positives. This suggests that the TSS proximity information could be used to filter out those false positives that exhibit the longest distances. Doing this effectively would however require more studying since a method for some distance threshold determination should be devised. Also, on short distances it is apparent that no major benefit is to be gained from this filtering scheme.

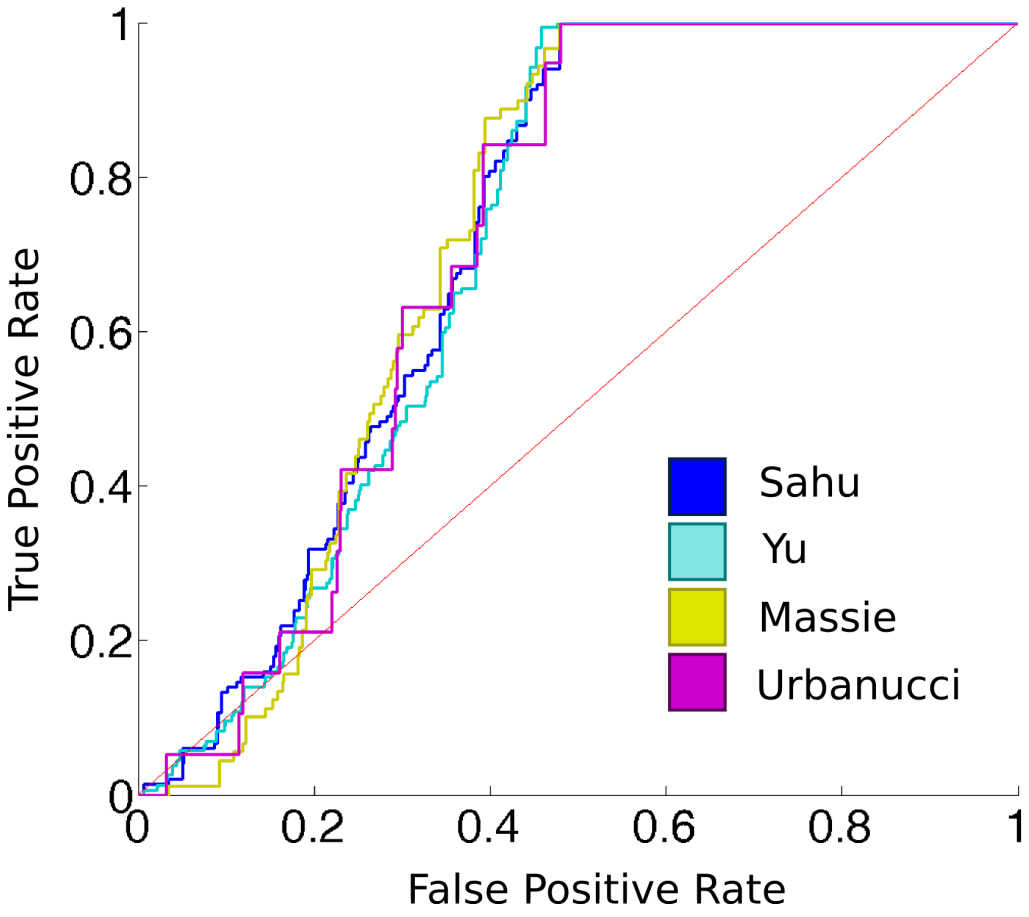

Figure 4.5: ROC curve of gene proximity thresholding

4.4 Predicted Network Inference

Based on the previous analyses in seemed that the performance of Housheng DNase-seq and ENCODE DNase-seq was quite similar. The Housheng data set seemed to detect slightly more ChIP-peaks and it was selected for visualization. First, a loosely constructed network including all candidate promoter and enhancer sites was inferred by stating that all predicted binding sites located 250kb upstream of a gene promoter are candidate regulators for the gene. The resulting network consisted of approximately 3.5 million regulation connections between TFs and genes, which on average translates into over $5,000$ regulated genes for each TF. When examining, for instance, the AR targets the predictions state that nearly $10,200$ genes are regulated by AR - an amount that seems rather excessive.

Also a more strict network construction with predicted regulation connections overlapping 1000bp sequences upstream of TSSs was also performed. The resulting network contained approximately 1.1 million connections between different regulators and targets. Even though the amount of connections drops the change is not dramatic. This is hyphotesized to be caused of the great number of false positive predictions showing regulation connections also there where they do not exist. Although principally fully visualizable, this network is still too large for a sensible visualization and the imaging of the networks was ignored here.

4.5 Gene Expression Enriched Networks

It appears that the prediction methods used have been able to produce a network of putative regulation connections in LNCaP cells given their current state. Clearly according to our ChIP-seq validations not all of these predicted connections are present in the cells making the network too ambiguous to be used as a stand-alone tool. To compare, validate and refine our predicted regulation network gene expression microarray data was used to enrich the constructed network.

The gene expression data created originally by Massie et al. consisted of a time course measurement of DHT treated LNCaP cells against ethanol treated controls. From the time series expression of steady state and 4 hour measurements were used aiming to uncover different waves of regulation and picture these as different layers in our gene regulation network. The data was downloaded from GEO [50] in a pre-processed form as a text file. The Illumina probe names were mapped to gene names and quantile normalization was performed to the data to ensure the validity between-chip comparisons. Differential expression was calculated based on at least two-fold difference in treated versus control samples and statistical significance shown by pvalue less than 0.05. [13]

Genes with log_2-scale expression under 6 in the steady state were removed and

the remaining genes were considered to be expressed. These 9419 genes were used to filter our predicted Millipede filtered network and as a result a network of 94888 regulation connections was obtained, reducing the number of connections substantially approximately to 8% of the original number. However, even this network is still much too large for sensible visualizations so network images were not produced.

Of the almost $100,000$ putative regulation connections of the previous network 1303 genes were claimed to be under AR regulation. The amount seems still quite large and still reflects the observed great numbers of false positives likely resulting into false connections. The differentially expressed genes at 4h in the gene expression measurement were extracted resulting in 23 genes which were all observed among the predicted targets of AR and thus among the 1303 genes mentioned above. The expressed targets of these genes were also obtained and the resulting two-level AR downstream network contained a total of 12984 regulation connections.

As this large network, although visualizable, cannot be presented in the text in a clear way, instead a very simple two-level network of predicted AR targets deemed differentially expressed by the microarray data analysis at 4 hour time point was illustrated. AR known target KLF5 [60] was identified from the predictions and from differentially expressed genes, and it was visualized with five of its predicted and differentially expressed targets to Figure 4.6. The genes are colored according to their differential expression, brighter red indicating larger difference. This visualization is a simple example of what kind of local network inference can be performed with the predicted network and gene expression data.

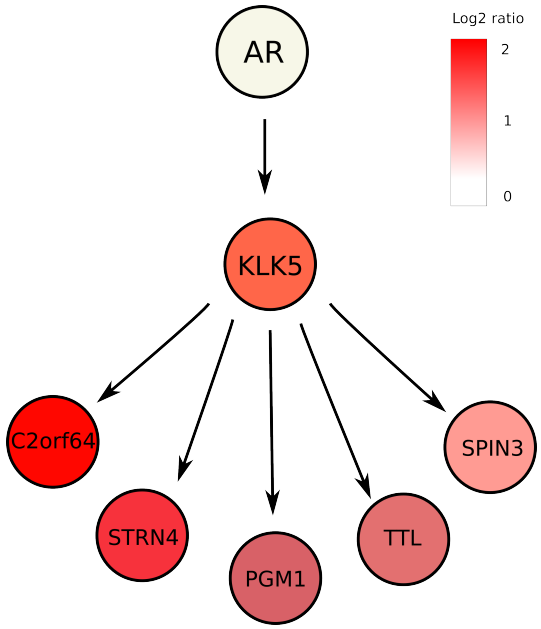

Figure 4.6: Two-level downstream regulation induced by AR

5. CONCLUSIONS

In this work computation prediction of TF binding using PWM scanning was merged to DNase I sequencing data in an attempt to create reliable predictions of regulation connections and, ultimately, a genome-wide gene regulatory network in LNCaP cells. The aim was achieved in the sense that all necessary tools needed for the process, what we could call 'de novo regulome assembly' from predictions, were either created or presented. The attempt was finally led to a conclusion by generation of gene expression information enriched global regulation network and AR downstream two-layer network.

The validation ChIP-seq datasets were shown to be solid performers throughout the work when compared to one another. Despite having very different amounts of peaks the sets showed mainly very consistent behaviour in ROC curves and in true positive predictions. The most notable exceptions were in validating the performance of DNase peak based PWM scans where 75%-80% of Sahu and Urbanucci datasets could be explained while only 46%-54% of Massie and Yu sets were predicted correctly. The differences in the case of the Yu set were thought to be caused by the different treatment of cells as they were cultivated in androgen-rich environment for several hours longer than in other experiments.

From the very first results with the naive DNase-seq utilization it became clear that the most challenging part of the approach is to find a way to effectively reduce the number of false positive predictions. Extensive efforts were made to overcome this challenge through various developed methods and new ideas. Compared to control ChIP-sequencing peaks the examined methods performed fairly, at their best explaining 75% of the ChIP all detectable peaks. However, despite these various efforts the false positive issue remained a dominating feature in all prediction types.

The reasons why the remaining portion of detectable peaks was not detected remains as a subject for speculations. It can very well be that some unfortunate patterns in DNase-seq signal or in the peaks detected from it gaps exist in crucial locations, preventing the detection. In future developments clearly the role of peak detection must be investigated further, it being one of the few aspects not covered in this thesis. As the peak detection affects directly the scanned sequences in other approaches than our naive starting point, different peak calling algorithms would surely have an impact on the prediction results.

Another interesting question is the proportion of ChIP-sequencing peaks that were deemed undetectable due to their location in areas of low PWM scores. Here most probably a balance of ChIP-sequencing biases, peak detection issues and effects of indirect regulation together form the observed behavior. Like all experiments, ChIP-seq is susceptible to many measurement errors, and the peak detection algorithms used, although nowadays very developed, have to in practice deal with noisy datasets of varying quality. Experiment and analysis related errors are thus always prone to exist and explain part of the undetectable peaks. Another major reason already shortly discussed in the results is the tendency of TFs to function in protein-protein interaction complexes. This leads to regulation control which does not require DNA binding from all subunits of the complex making sites these subunits impossible to detect at these locations with DNA sequence scanning methods. However, as ChIP-seq experiment may very well detect such complexes and corresponding sequence is reported as response element to also those proteins not actually binding the DNA. The inability to take into account indirect regulation is a clear weakness of all DNA sequence scanning based methods with no apparent solution.

Keeping the aforementioned in mind it must be emphasized here that the reasonability of the performed validations depends critically on the available ChIP-seq dataset and all the reported results also in this thesis are valid only so far as the ChIP-seq data can be judged trustworthy. This means that dividing the prediction results directly into 'right' and 'wrong' or true and false predictions is somewhat unrealistic as the cell as a dynamic environment is a subject to constant variations and the regulation patterns may quickly be altered. The so called ChIP-seq ground-truth is always a snapshot of the cell in one condition and is thus never alone capable to picture the all regulation connections even in same cell type and similar conditions. ChIP-seq can, however, be considered as a rather trustworthy method in determining at least what connections are present making it a fine tool for validating our predictions as long as the here mentioned limitations are kept in mind when interpreting the results.

Some simplifications had to be made in the method construction and validations and these need to be given more attention in future developments. The selection of initial filtering criterion for PWM scan scores is somewhat arbitrary since no guidelines and no apparent testing method existed at the point when the choice had to be made. As noticed in the validations, however, the original seemingly strict filtering criterion seemed actually to have been a quite reasonable choice, as lowering the criterion was shown to explode the number of false positives without clear benefits in detection. Another clear simplification is the excessive use of AR in PWM scan score threshold assessment and in all validations. This special role for AR in this thesis was partially forced and partially wanted choice. As a key regulator

in prostate cancer AR is an interesting target for the information revealed by the validations and, later, the gene regulatory network construction. Also since AR has been a subject for intensive study a multitude of datasets mapping its characteristics exists and are freely available. This largely enabled the broad inspections presented in this thesis in the context of different ChIP-seq set comparisons, DNase-seq data comparisons and utilization of time course gene expression data. This positive factor is in the same time the reason for forced selection of AR as our main target molecule: not as excessive datasets exist for other factors.

According to earlier published work the Millipede and Centipede produce highly more accurate results than the methods reviewed in this work. However, these methods typically utilize PWM scanning and DNase-seq signals but require also ChIP-seq inputs for best performance our approach has certain clear advantages. Since ChIP-sequencing data is not available for even nearly all transcription factors in all common cell types the methods utilizing it are futile when it comes to whole genome-wide regulome assembly. This is why further development of our approach is potentially very interesting due to its general nature.

The future prospects of these types of approaches attempting prediction based regulome assembly is somewhat open. While potential applications of reliable genome-wide regulation networks in human cells are intriguing, the challenges on the way in achieving this goal are still considerable. However, we can fairly comfortably state that such attempts will be seen in the future more often and hopefully also with developed methods. Still, to which extent this direction of research is pursued remains a mystery.

The presented work is a thorough exploration of steps required for prediction based regulome assembly. While producing fair results perhaps its most value lies in the knowledge of approaches that do not result improvements in detection precision so that these can be avoided in the future. Altogether the thesis forms a solid foundation for future studies aiming to pursue even more ambitious goals in the field of computational prediction of regulation.

REFERENCES

[1] Word Cancer Reserach Fund International (2014, Mar. 4), *Worldwide Cancer Statistics* [Online], Available at: http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php

[2] David L. Nelson and Michael M. Cox "Lehninger: Principles of Biochemistry", 5th ed. New York: W.H. Freeman and Company , 2008

[3] Andrew J Bannister and Tony Kouzarides, "Regulation of chromatin by histone modifications", *Cell Research*, 21(3): 381-395, March 2011

[4] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter, "Essential Cell Biology", 3th ed. New York: Garland Science, 2010

[5] Lodish H, Berk A, Zipursky SL, "Molecular Cell Biology" 4th ed. New York: W. H. Freeman, 2000. Section 11.2, Processing of Eukaryotic mRNA.

[6] Amy Ralston, *Gene Expression Regulates Cell Differentiation*, [Online]. Available: http://www.nature.com/scitable/topicpage/gene-expression-regulates-cell-differentiation-931

[7] Gary D. Stormo, "DNA binding sites: representation and discovery", *Bioinformatics*, Vol. 16 no. 1, pp. 16-23, 2000

[8] Francois Spitz and Eileen E. M. Furlong, "Transcription factors: from enhancer binding to developmental control", Nature Reviews Genetics 13, 613-626, 2012

[9] Nicolas M. Luscombe, Madan M. Babu, Haiyuan Yu, Michael Snyder, Sarah A. Teichmann, Mark Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes", *Nature* Sep. 16;431(7006):308-12, 2004

[10] Tom Quick, Chrystopher L. Nehaniv, Kerstin Dautenhahn, Graham Roberts, "Evolving Embodied Genetic Regulatory Network-Driven Control Systems", *Lecture Notes in Computer Science* Volume 2801, pp 266-277, 2003

[11] Eric H. Davidson, "Emerging properties of animal gene regulatory networks", *Nature* 468, pp. 911–920, 2010

[12] Zoulfia Darieva, Anne Clancy, Richard Bulmer, Emma Williams, Aline Pic-Taylor, Brian A. Morgan, Andrew D. Sharrocksemail, "A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression", *Molecular Cell*, Apr 9;38(1), pp. 29-40, 2010

[13] Charles E Massie, Andy Lynch, Antonio Ramos Montoya, Joan Boren, Rory Stark, Ladan Fazli, Anne Warren, Helen Scott, Basetti Madhu, Naomi Sharma, Helene Bon, Vinny Zecchini, Donna Michelle Smith, Gina M DeNicola, Nik Mathews, Michelle Osborne, James Hadfield, Stewart MacArthur, Boris Adryan, Scott K Lyons, Kevin M Brindle, John Griffiths, Martin E Gleave, Paul S Rennie, David E Neal, Ian G Mills, "The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis", *The EMBO Journal*, May 30(13), pp. 2719 2733, 2011

[14] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin RAff, Keith Roberts, Peter Walter, "Molecular Biology of the Cell", 5th ed. New York: Garland Science, 2008

[15] Elsevier Clinical Key, *Benign Prostatic Hyperplasia* [ONLINE], Available: https://www.clinicalkey.com/topics/urology/benign-prostatic-hyperplasia.html

[16] American Cancer Society, *Key Statistics of Prostate Cancer* [Online], Available: http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-key-statistics

[17] Brian J. Feldman and David Feldman, "The development of androgen-independent prostate cancer", *Nature Reviews Cancer* 1, pp. 34-45, 2001

[18] James L. Mohler, Christopher W. Gregory, O. Harris Ford III, Desok Kim , Catharina M. Weaver, Peter Petrusz, Elizabeth M. Wilson, Frank S. French, "The Androgen Axis in Recurrent Prostate Cancer", *Clinical Cancer Research*, 10, pp. 440-448, 2004

[19] Christopher W. Gregory, Bin He, Raymond T. Johnson, O. Harris Ford, James L. Mohler, Frank S. French, and Elizabeth M. Wilson, "A Mechanism for Androgen Receptor-mediated Prostate Cancer Recurrence after Androgen Deprivation Therapy", *Cancer Research*, 61(11), pp: 4315-4319, 2001

[20] "Cell lines used in prostate cancer research: a compendium of old and new lines–part 1.", *Journal of Urology*, 173(2), pp. 342-359, 2005

[21] Pamela J. Russell and Elizabeth A. Kingsley, "Human Prostate Cancer Cell Lines", *Prostate Cancer Methods and Protocols*, XII, pp. 403, 2003

[22] Frederick Sanger and A R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase", Vol. 94 Issue 3, pp 441-446, 1975

[23] Michael L. Metzker, "Emerging technologies in DNA sequencing", *Genome Research*, 15, pp. 1767-1776, 2005

[24] Michael L. Metzker, "Sequencing technologies — the next generation", *Nature Reviews Genetics*, 11(1), pp. 31-46, 2010

[25] Schraga Schwartz, Ram Oren, Gil Ast, "Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads", *PLOS ONE* 31. Jan., 2011

[26] Illumina, *Technology spotlight: Illumina Sequencing* [Online], Available: http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pd

[27] Lingyun Song and Gregory E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells", *Cold Spring Harbor Protocols*, Feb., 2010

[28] Jeremy M. Simon, Paul G. Giresi, Ian J. Davis, Jason D. Lieb, "A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE)", *Current Protocols in Molecular Biology*, Chapter 21 Unit 21, 2013http://www.ncbi.nlm.nih.gov/pubmed/23547014

[29] Peter J. Park, "ChIP–seq: advantages and challenges of a maturing technology", *Nature Reviews Genetics*, 10, pp. 669-680, 2009

[30] David S. Johnson, Ali Mortazavi, Richard M. Myers, Barbara Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions", *Science*, 8, 2007

[31] Jörg D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis", *Nature Reviews Genetics*, 7, pp. 200-210, 2006

[32] Melissa B. Miller and Yi-Wei Tang, 'Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology, *Clinical Microbiology Reviews*, 22(4), pp. 611, 2009

[33] Illumina, *BeadArray Microarray Technology* [Online], Available: http://www.illumina.com/technology/beadarray_technology.ilmn

[34] Leonid Teytelmana, Deborah M. Thurtlec, Jasper Rinec, and Alexander van Oudenaardena, "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110 no. 46, pp. 18602-18607, 2013

[35] Housheng Hansen He, Clifford A Meyer, Sheng'en Shawn Hu, Mei-Wei Chen, Chongzhi Zang, Yin Liu, Prakash K Rao, Teng Fei, Han Xu, Henry Long, X Shirley Liu , Myles Brown, "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification", *Nature Methods*, 11, pp. 73-78, 2014

[36] Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome Biology*, 10, 2005

[37] Cole Trapnell1, Lior Pachter and Steven L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq", *Bioinformatics*, Vol. 25 Issue 9, pp. 1105-1111, 2009

[38] Heng Li, Jue Ruan, and Richard Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome Reserach*, 18, pp. 1851-1858, 2008 http://genome.cshlp.org/content/18/11/1851.long

[39] Yong Zhang, Tao Liu, Clifford A Meyer, Jerome Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li and X Shirley Liu, "Model-based analysis of ChIP-Seq (MACS)", *Genome Biology*, 9(9), pp. R137, 2008

[40] Sven Heinz, Christopher Benner and Christopher K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities", *Molecular Cell*, 38(4), pp. 576-589, 2010

[41] Nomenclature Committee of the International Union of Biochemistry, *Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences* [Online], Available: http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html

[42] Xuhua Xia, "Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction", *Scientifica*, Vol. 2012, 2012

[43] Kaixuan Luo and Alexander J. Hartemink, "Using DNase digestion data to accurately identify transcription factor binding sites", *Pacific Symposium on Biocomputing*, pp. 80-91, 2013

[44] Roger Pique-Regi Jacob F. Degner Athma A. Pai, Daniel J. Gaffney, Yoav Gilad and Jonathan K. Pritchard, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data", *Genome Research*, 21(3), pp. 447-477, 2011

[45] Housheng Hansen He, Clifford A. Meyer and X. Shirley Liu, "Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics", *Genome Research*, 22(6), pp. 1015-1025, 2012

[46] Gene Expression Omnibus, *Fine mapping of androgen regulated genes in LNCaP cells* [Online], Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM822388

[47] Biswajyoti Sahu, Marko Laakso, Päivi Pihlajamaa, Kristian Ovaska, Ievgenii Sinielnikov, Sampsa Hautaniemi, and Olli A. Jänne, "FoxA1 specifies unique androgen and glucocorticoid receptor binding events in prostate cancer cells", *Cancer Research*, 73(5), pp. 1570-1580, 2013

[48] Jindan Yu, Jianjun Yu, Ram-Shankar Mani, Qi Cao, Chad J. Brenner, Xuhong Cao, Xiaoju Wang, Longtao Wu, James Li, Ming Hu, Yusong Gong, Hong Cheng, Bharathi Laxman, Adaikkalam Vellaichamy, Sunita Shankar, Yong Li, Saravana M. Dhanasekaran, Roger Morey, Terrence Barrette, Robert J. Lonigro, Scott A. Tomlins, Sooryanarayana Varambally, Zhaohui S. Qin, Arul M. Chinnaiyan, "An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression", *Cancer Cell*, 17(5), pp. 443-454, 2010

[49] Urbanucci A, Sahu B, Seppälä J, Larjo A, Latonen LM, Waltering KK, Tammela TL, Vessella RL, Lähdesmäki H, Jänne OA, Visakorpi T, "Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer", *Oncogene*, 31(17), 2153-2163, 2012

[50] Gene Expression Onmibus, *Fine mapping of androgen regulated genes in LNCaP cells* [Online], Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18684

[51] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, François Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin2 and Wyeth W. Wasserman, "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles", *Nucleic Acid Research*, Vol. 42 Issue D1 ,pp. D142-D147, 2013

[52] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman and Boris Lenhard, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles", *Nucleic Acid Research*, Vol. 32 Issue suppl. 1, pp. D91-D94, 2004

[53] Michael F. Berger and Martha L. Bulyk, "Universal protein binding microarrays for the comprehensive characterization of the DNA binding specificities of transcription factors", *Nature Protocols*, 4(3), pp. 393-411, 2009

[54] UniPROBE, *UniPROBE Database* [Online], Available: http://the_brain.bwh.harvard.edu/uniprobe/

[55] Daniel E. Newburger and Martha L. Bulyk, "UniPROBE: an online database of protein binding microarray data on protein-DNA interactions", *Nucleic Acids Research*, 37, pp. D77-D82, 2009

[56] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender, "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes", *Nucleic Acid Research*, Vol. 34 Issue suppl1, pp. D108-D110, 2005

[57] Biobase Biological Databases, *TRANSFAC Professional*, [Online], Available: http://www.gene-regulation.com/index2.html

[58] TRANSFAC, *TRANSFAC Professional vs Public* [Online], Available: https://portal.biobase-international.com/archive/documents/transfac_comparison.pdf

[59] H Makkonen, T Jääskeläinen, T Pitkänen-Arsiola, M Rytinki, K K Waltering, M Mättö, T Visakorpi and J J Palvimo, "Identification of ETS-like transcription factor 4 as a novel androgen receptor target in prostate cancer cells", *Oncogene*, 27(36), pp. 4865-4876, 2008

[60] Daniel E. Frigo, Andrea B. Sherk, Bryan M. Wittmann, John D. Norris, Qianben Wang, James D. Joseph, Aidan P. Toner, Myles Brown, and Donald P. McDonnell, "Induction of Kruppel-like factor 5 expression by androgens results in increased CXCR4-dependent migration of prostate cancer cells in vitro", *Molecular Endocrinology*, 23(9), pp. 1385-1396, 2009